Speech

Multilingual Speech Recognition and Synthesis

Feb. 27, 2024

Columbia University

Presenter: Bhuvana Ramabhadran

Contributors: Speech/Research Teams @ Google

Outline

- Challenges of multilingualism
- ASR
  - Scaling to many languages
  - Taking advantage of found data
- TTS
  - Can we share the same ideas from ASR?
  - Understanding shared representations of multiple modalities key?
- Concluding thoughts

Multilingual models

- State-of-the-art
  - Allow for joint training of data-rich and data-scarce languages in a single model
  - Require the encoding of language information which makes it less flexible

- Can we build a language-agnostic multilingual ASR system?
  - Challenge: Can similar sounding acoustics across languages be mapped to a single, canonical target sequence of graphemes or sub-word units?

Desirable Characteristics of multilingual models

- **Modeling/Systems techniques**
    - Language Expansion
        - Enables using the same model with/without knowing language-id
        - Language ID decision made using the same ASR encoder
    - Model Capacity
    - Decoder allows for Parallel-beam search
    - Applications to multiple tasks
        - Speech Recognition, Translation, Synthesis, etc.
- **Production considerations**
        - ASR is more than just the e2e model
        - Recognition cost / Quality / Maintainability / Refresh

Prior work

- Prior work in training multilingual representations [1, 2] and end-to-end models [3, 4] have demonstrated that the best performing models require conditioning on language information

    - [1] B. Ma, C. Guan, H. Li, and C.-H. Lee, “Multilingual speech recognition with language identification,” in ICSLP, 2002.
    - [2] A. Cutler, Y. Zhang, E. Chuangsuwanich, and J.R. Glass, “Language ID-based training of multilingual stacked bottleneck features,” in Interspeech, 2014.
    - [3] S. Watanabe, T. Hori, and J.R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in ASRU, 2017.
    - [4] A. Kannan, A. Datta, T.N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, “Large-scale multilingual speech recognition with a streaming end-to-end model,” arXiv preprint arXiv:1909.05330, 2019.

Prior work (contd.)

- Need to track language switches within an utterance [5, 6], adjust language sampling ratios, or add additional parameters based on the data distribution [4]

  - [5] H. Seki, S. Watanabe, T. Hori, J. Le Roux, and J.R. Hershey, “An end-to-end language-tracking speech recognizer for mixed-language speech,” in ICASSP, 2018.
  - [6] A. Waters, N. Gaur, P. Haghani, P. Moreno, and Z. Qu, “Leveraging language id in multilingual end-to-end speech recognition,” in ASRU, 2019.

What language clusters and why?

For example, many Indic languages can we cover with one model?

- Take advantage of overlap in acoustic and lexical content
    - due to either language family relations or the geographic and cultural proximity of the native speakers.
- However, their writing systems occupy different unicode blocks
- Can we combine languages from multiple languages families efficiently and produce “usable” models for users?

...what challenge does this pose?

Challenges: Code-Switching

- Code-switching is a commonly occurring phenomenon in many multilingual communities, wherein a speaker switches between languages within a single utterance (Hindi-English, Bengali-English, Arabic-English and Chinese-English, Spanish-English, etc.)
- Can occur at morphological, lexical, syntactic, semantic, pragmatic levels
- A good read on Bilingual Speech from a linguistic perspective:
  - Analysis of many language-pairs
  - Bilingual verbs: the phenomenon of verbal compounds combining elements from two languages
  - Impact of psycholinguistic and social factors : language dominance, duration of contact, bilingual proficiency, speaker type, age-group or generation and language attitudes.

Pieter Muysken, Bilingual speech: A typology of code-mixing. Cambridge: Cambridge University Press, 2000.

Examples of code-switching

- Words with different language indices are inserted into a phrase structure
- Spanish-English
  - Cuando mi novio *tweet*ea pero no contesta (When my boyfriend tweets but doesn’t answer)
  - Agarrar *my Master’s* (Get my Master’s)
- Ambiguities in transcription
  - *डिस्कवरी vs discovery*
  - *होम्योपथी में अर्थराइटिस treatment vs Homeopathy में arthritis treatment*
- These *rendering* errors artificially inflate the **W**ord **E**rror **R**ate (WER)
- Harder to differentiate between ***modeling*** and ***rendering*** errors
  - *fancy साड़ी दिखाइए vs fancy Sadi dikhaiye*

Code-Switching

- Handled the problem of foreign word pronunciation using language dependent phonemes by creating linguistically motivated pairwise mappings for each language involved in code-switching.

  White, Christopher M., Sanjeev Khudanpur, and James K. Baker. "An investigation of acoustic models for multilingual code-switching." *Ninth Annual Conference of the International Speech Communication Association*. 2008.

- In Mandarin-English use of combined subwords from both languages as modeling units along with an additional objective of training with language ID was found to be useful.

  Luo, Ne, et al. "Towards end-to-end code-switching speech recognition." *arXiv preprint arXiv:1810.13091* (2018).

Code-Switching

- Separately train an E2E CTC model and a frame-level language identification (LID) model. Linearly adjust the posteriors of an E2E CTC model using the LID scores (Mandarin-English)

  Li, Ke, et al. "Towards code-switching ASR for end-to-end CTC models." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

- Effectiveness of multilingual models on NLU tasks such as named entity recognition and part-of-speech tagging tasks (Hindi-English, Spanish-English, and Modern Standard Arabic-Egyptian)? Pretrained multilingual models not as effective as hierarchical embeddings to deal with code-switching

  White, Christopher M., Sanjeev Khudanpur, and James K. Baker. "An investigation of acoustic models for multilingual code-switching." Ninth Annual Conference of the International Speech Communication Association. 2008.

Code-Switching

- In Frisian-Dutch merging phones of both languages provides the best recognition performance for code-switched words

  Yılmaz, Emre, Henk van den Heuvel, and David Van Leeuwen. "Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech." Procedia Computer Science 81 (2016): 159-166.

- Data Augmentation by generating synthetic code-switched data with word translation or word insertion followed by audio splicing using text-to-speech

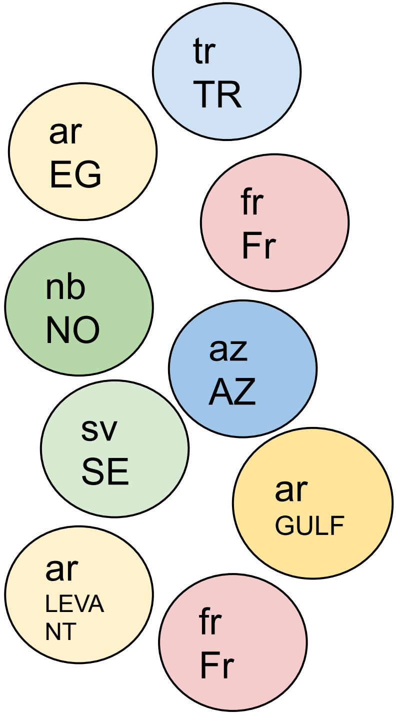
  Du, Chenpeng, et al. "Data augmentation for end-to-end code-switching speech recognition." *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021.

Code-Switching

- Output token embeddings of two monolingual languages are differently distributed; Constrain with Jensen-Shannon divergence to force embeddings of monolingual languages to possess similar distributions

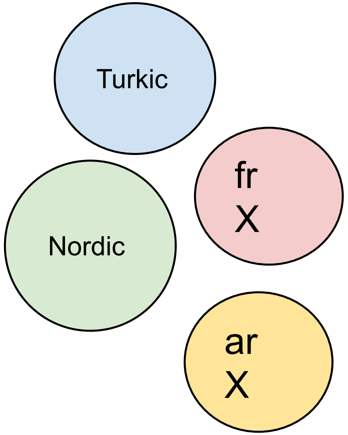
Khassanov, Yerbolat, et al. "Constrained output embeddings for end-to-end code-switching speech recognition with only monolingual data." *arXiv preprint arXiv:1904.03802* (2019).

Multilingual Model Clusters

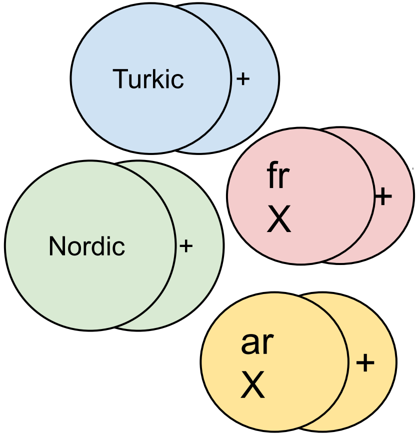


Previously

Multi-Language Clusters



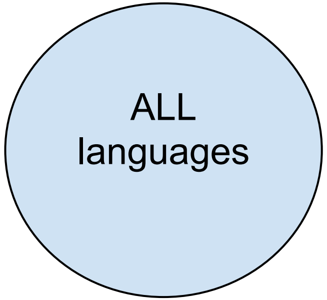
Reduce cost of multilingual recognition



Clusters augmented with each cluster's most commonly paired up language(s)

More common

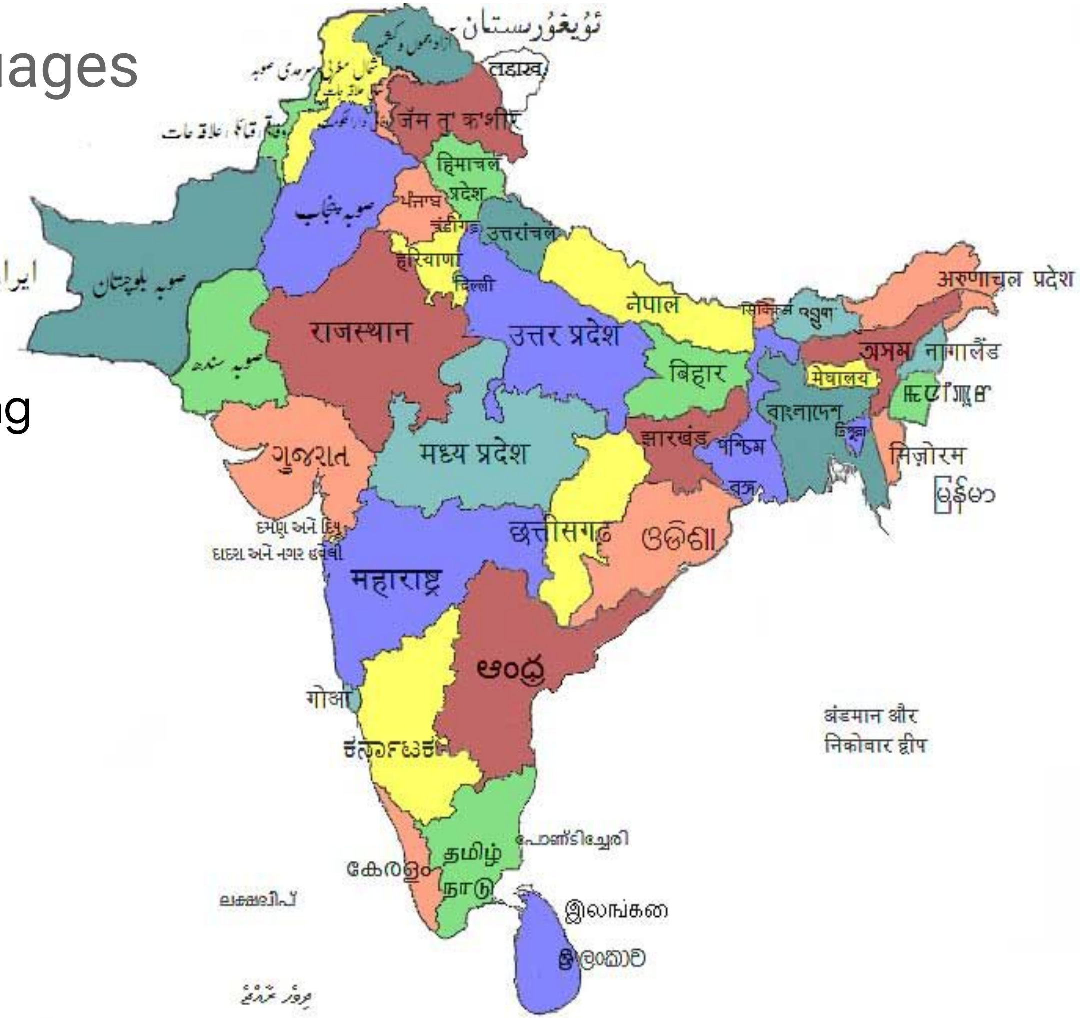
Towards one large model



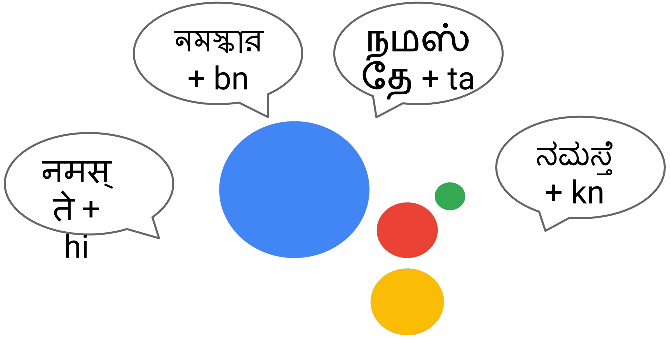
Now

South Asia: A land of languages

- **Scripts**: Several writing systems
- **Code-switching:** language mixing
- **Several** languages and dialects!
- But **overlap** due to linguistic similarity, and/or geographic & cultural proximity of the native speakers.



Language-dependent vs Language-agnostic



State-of-the-art performance
Language-id used to

➔ track language switches

➔ adjust language sampling ratios

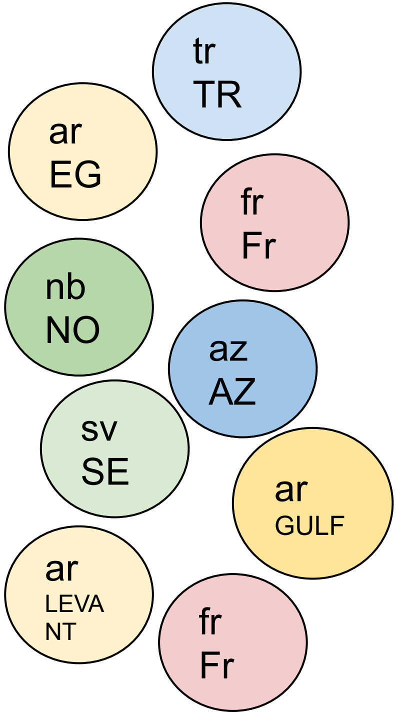
➔ add params based on data distribution

Transliterate all languages to same script (e.g. Latin).

Naturally handles code-switching.

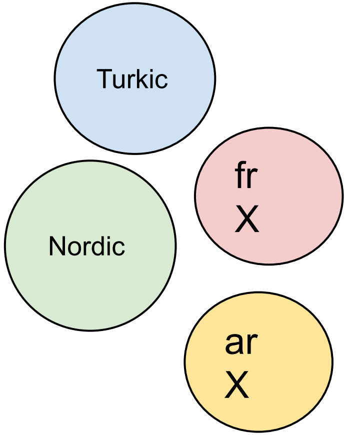
Easily scale to new/unseen languages.

Multilingual Model Clusters



Previously

Multi-Language Clusters



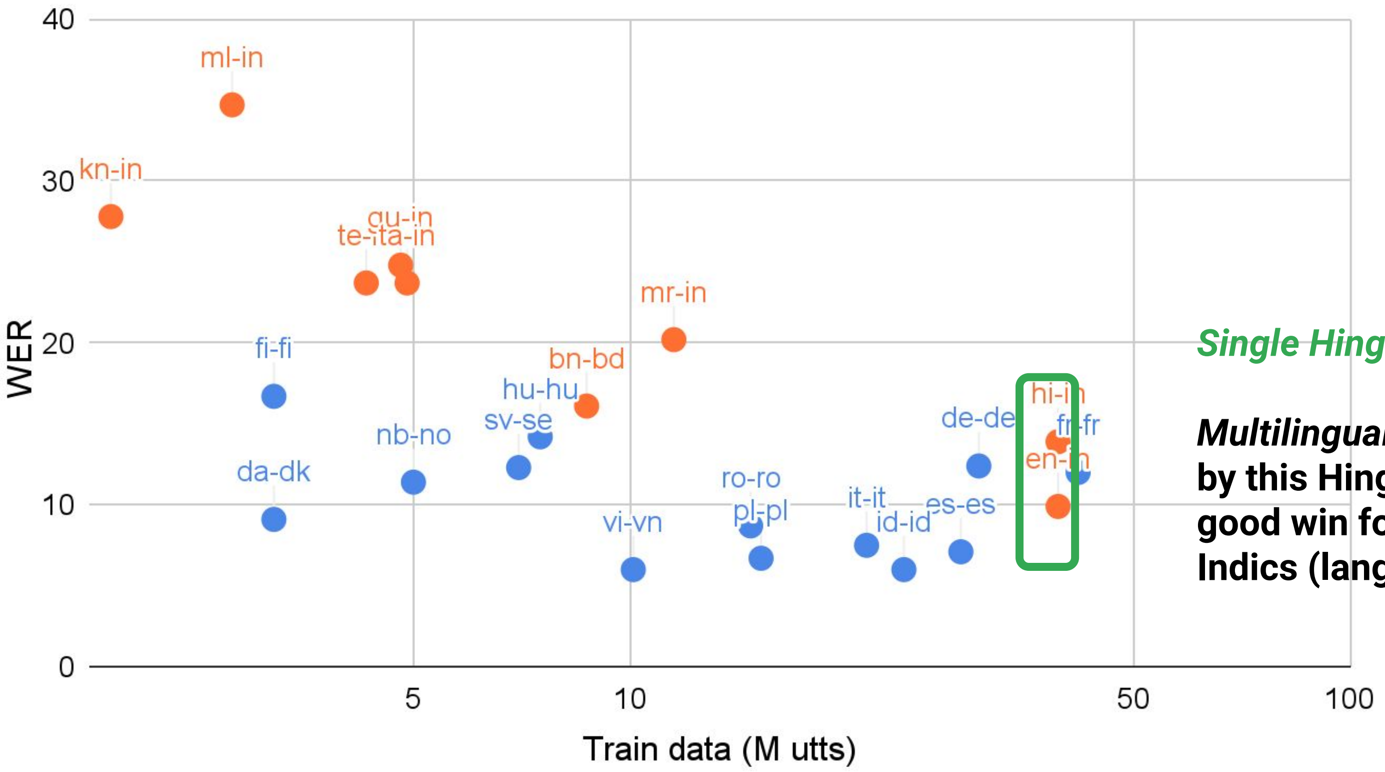
Today

ASR models for *clusters* of low-resource/related languages

- **Improve** lower-resource languages **performance** by pulling data from other languages
- Significantly **reduce** the training and inference **cost**

Language		WER	Mono Delta
Nordic	da_dk	12%	-0.22%
	nb_no	13%	-9.80%
	sv_se	12%	-18%
Arabic	ar_eg	13%	-17%
	ar_x_levant	14%	-18%
	ar_x_gulf	11%	-26%
	ar_x_maghrebi	11%	-11%
South Asian	kn_in	28%	-6.30%
	gu_in	20%	-28%
	mr_in	23%	1.00%
	te_in	26%	-6.50%
	si_lk	35%	-2.90%
	ur_pk	16%	-25%

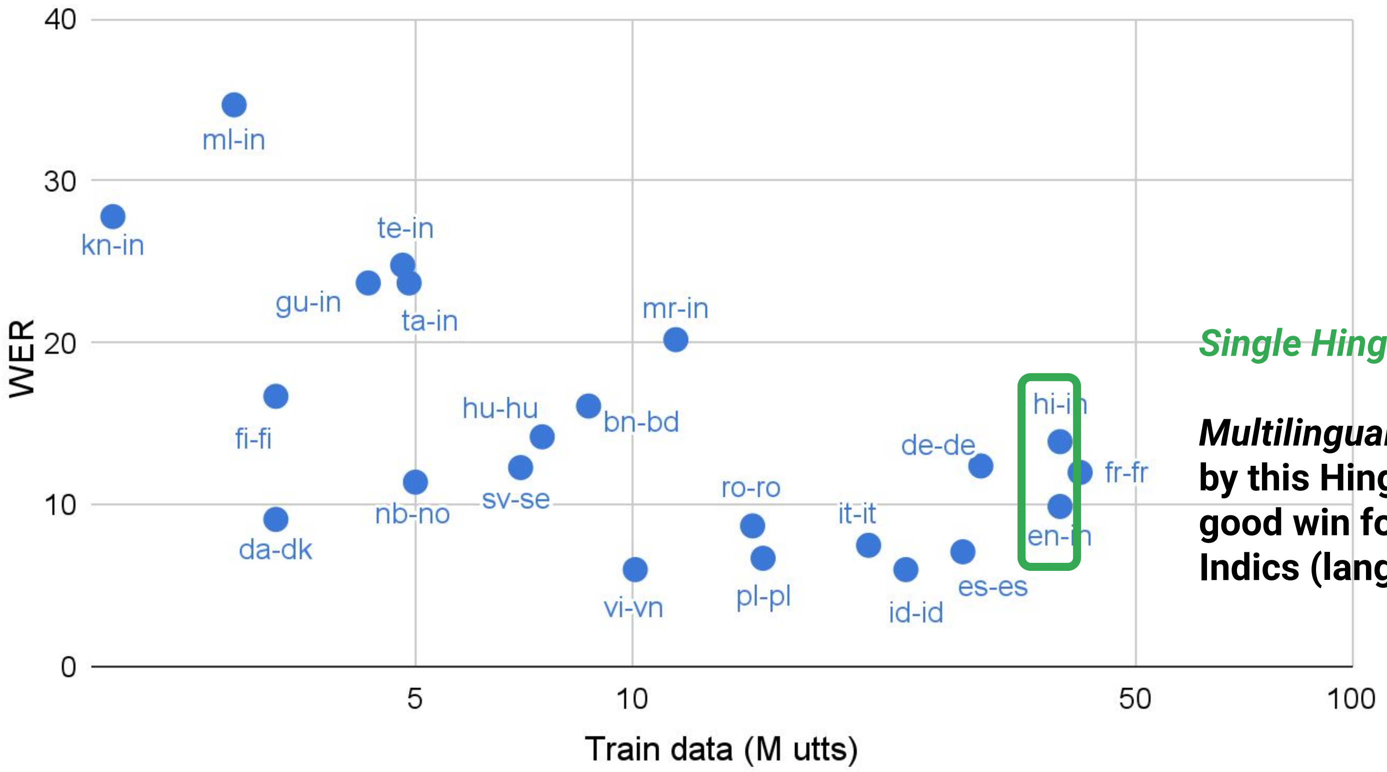
Data Distribution across languages



Single Hinglish model

***Multilingualism*: Initialization by this Hinglish model is a good win for monolingual Indics (language transfer)**

Data Distribution across languages



Single Hinglish model

***Multilingualism*: Initialization by this Hinglish model is a good win for monolingual Indics (language transfer)**

Challenge in multilingual transliteration

Attested romanizations of the English word “discovery”

Bengali ডিসকভারি	Hindi डिस्कवरी	Kannada ಡಿಸ್ಕವರಿ	Tamil டிஸ்கவரி
discoveri	discovery	discovary	tiskavari
discovery		discovery	discovery
diskovary		discoveri	
diskovery		discowery	
diskoveri			

Code-Switching Benchmark: For NLP research (https://ritual.uh.edu/lince/)

LinCE is a continuous effort, and we will expand it with more low-resource languages and tasks.

Language Pairs	LID	POS	NER	SA	MT
Spanish-English	✓	✓	✓	✓	
Hindi-English	✓	✓	✓		
Nepali-English	✓				
Modern Standard Arabic-Egyptian Arabic	✓		✓		
English-Hinglish					✓
Spanglish-English					✓
English-Spanglish					✓
(Modern Standard Arabic-Egyptian Arabic)-English					✓
English-(Modern Standard Arabic-Egyptian Arabic)					✓

Thoughts

- *Data Preprocessing:* Simple but effective approach to building language-agnostic multilingual ASR systems for Indic languages (we have explored up to 12 Indic languages today)
- *Parameter Sharing*: Reduced number of modeling units resulting from the use of one canonical writing system (Latin) allows to reduce ambiguities and build competitive multilingual models without language-ID
- *Data Balancing* for efficient knowledge transfer. Languages compete for capacity.Data-scarce language is overfitting while data-rich languages have not converged
- Currently, multilingual models show a performance gap with the best possible monolingual models in a production setting

How many languages can state-of-the-art technology handle?

Google

Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, Yonghui Wu

March 2023

State-of-the-art performance in ASR and ST tasks

- Efficient Pre-training
- Incorporating Untranscribed Speech, Unspoken Text, Paired Speech-Text
- Modality matching for in the Injection of unspoken text
- Language-ID
- Code-Switching

Bharadwaj, S., Ma, M., Vashishth, S., Bapna, A., Ganapathy, S., Axelrod, V., Dalmia, S., Han, W., Zhang, Y., van Esch, D. and Ritchie, S., 2023. Multimodal Modeling For Spoken Language Identification. arXiv preprint arXiv:2309.10567.

Google Universal Speech Model for 100+ Languages

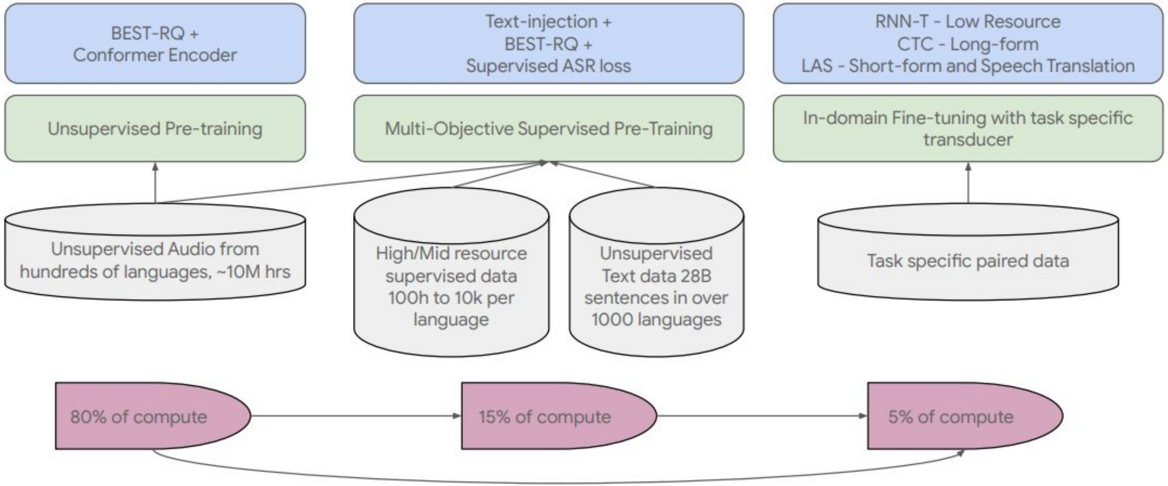


Figure 1: An overview of our approach. Training is split into three stages. (i) The first stage trains a conformer backbone on a large unlabeled speech dataset, optimizing for the BEST-RQ objective. (ii) We continue training this speech representation learning model while optimizing for multiple objectives, the BEST-RQ objective on unlabeled speech, the modality matching, supervised ASR and duration modeling losses on paired speech and transcript data and the text reconstruction objective with an RNN-T decoder on unlabeled text. (iii) The third stage fine-tunes this pre-trained encoder on the ASR or AST tasks.

Pretraining

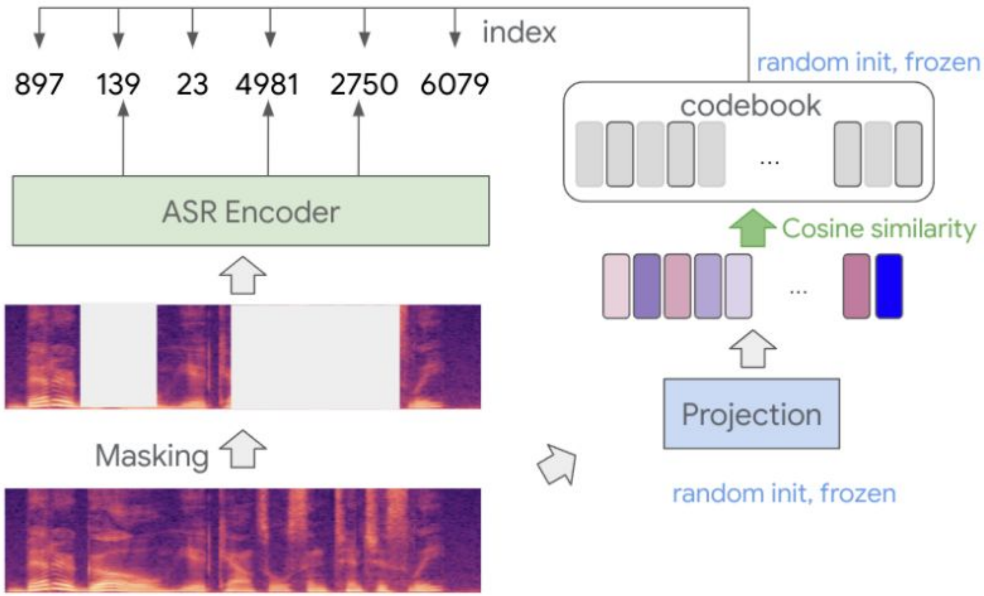


Figure 3: BEST-RQ based pre-training with conformer encoder.

BEST-RQ (BERT-based Speech pre-Training with Random projection Quantizer) is used to pre-train the encoder of the model 2B conformer

Text-Injection and modality matching

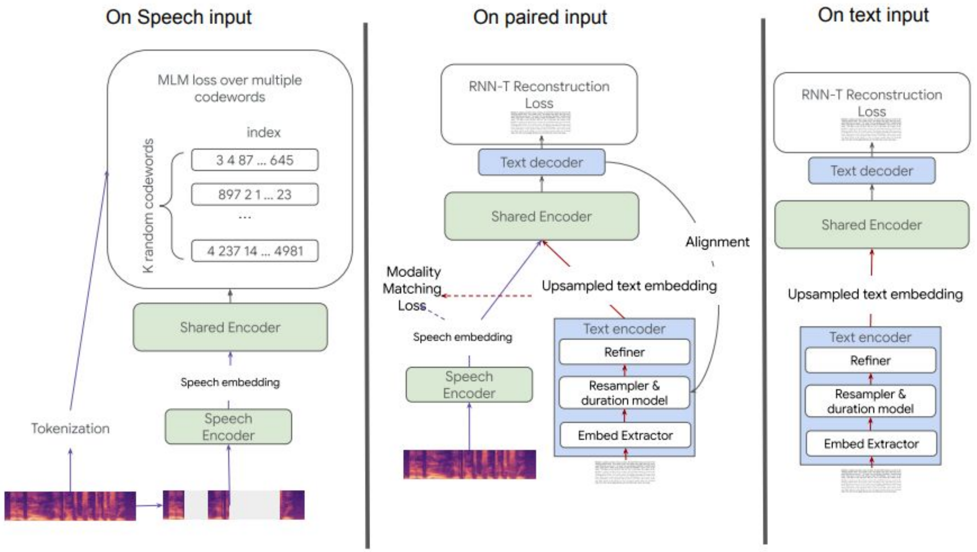


Figure 5: Overview of MOST text injection. The left-most panel depicts MOST training on unlabeled speech input; the center panel depicts training on paired speech and text input; the right-most panel depicts training on unlabeled text data.

Training Data

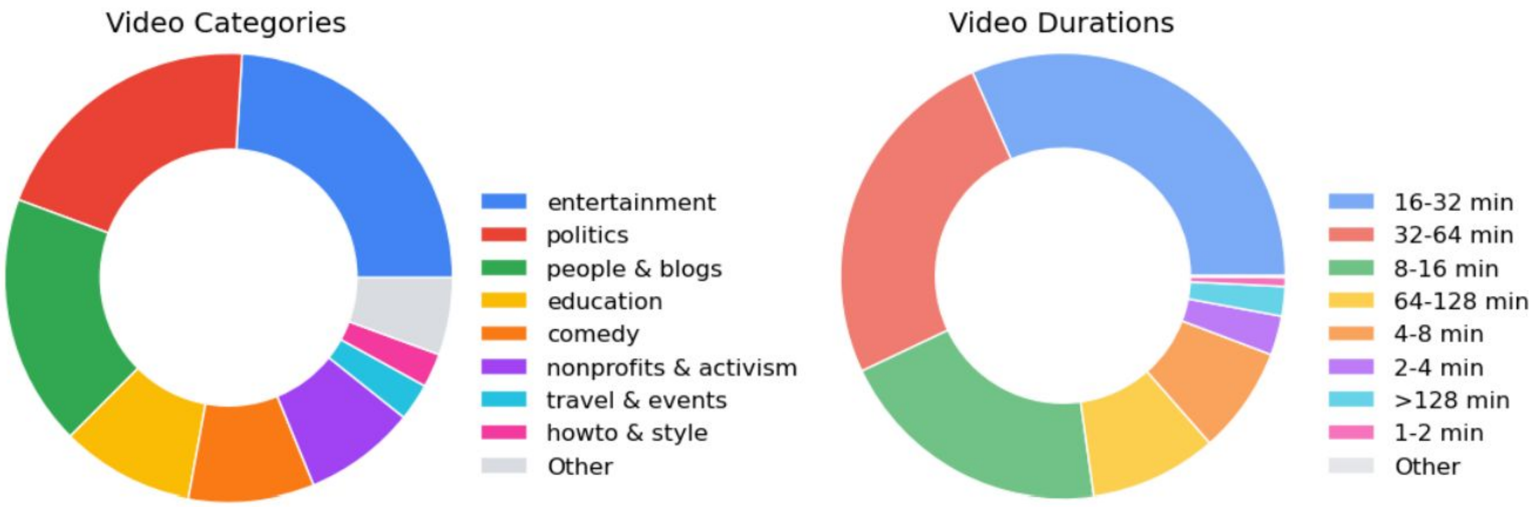


Figure 6: The video category and length distribution of YT-513-U.

Key Findings

- BEST-RQ is a scalable speech representation learner: We find that BEST-RQ pre-training can effectively scale to the very large data regime with a 2B parameter Conformer-based backbone.
- MOST (BEST-RQ + text-injection) is a scalable speech and text representation learner: It is an effective method for utilizing large scale text data for improving quality on downstream speech tasks, as demonstrated by quality gains exhibited for the FLEURS and CoVoST 2 tasks.
- Representations from MOST (BEST-RQ + text-injection) can quickly adapt to new domains with light-weight residual adapters.
- SoTA results for downstream multilingual speech tasks:
    - SpeechStew (mono-lingual ASR)
    - CORAAL (African American Vernacular English (AAVE) ASR)
    - FLEURS (multi-lingual ASR) [16], YT (multilingual long-form ASR)
    - CoVoST (AST from English to multiple languages).

Scalability: Language Expansion Results

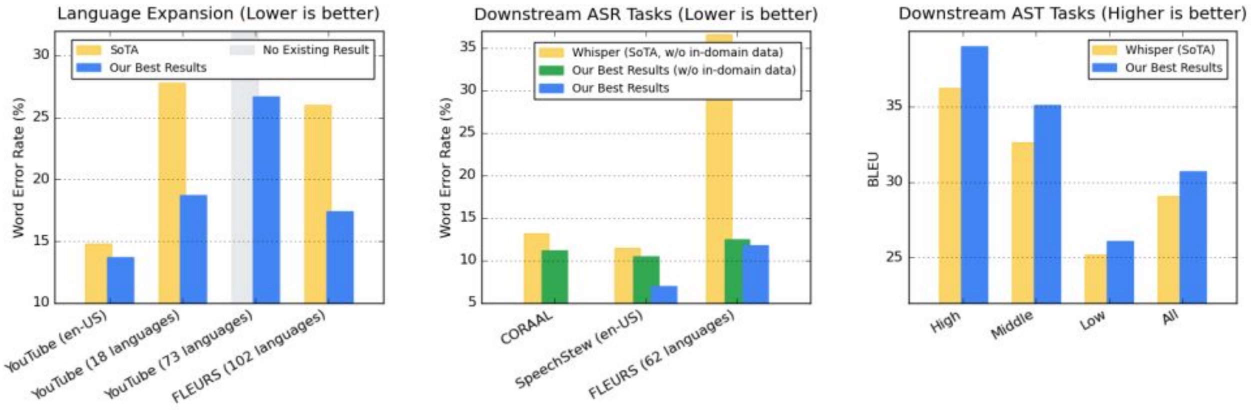
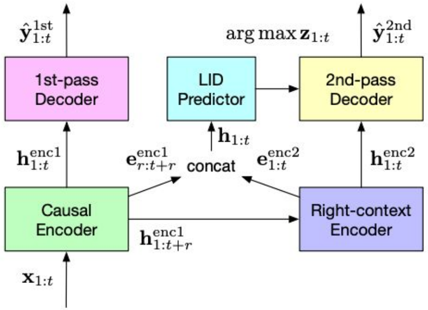


Figure 2: **(Left)**† WERs (%) Our language expansion effort to support more languages on YouTube (73 languages) and extending to 100+ languages on the public dataset (FLEURS). Lower is better. To the best of our knowledge, no published model can successfully decode all 73 languages from our YouTube set, thus we only list our results. **(Middle)**† Our results on ASR benchmarks, with or without in-domain data. Lower is better. **(Right)** SoTA results on public speech translation tasks. Results presented are presented as high/middle/low resources languages defined in [20]. Higher is better.

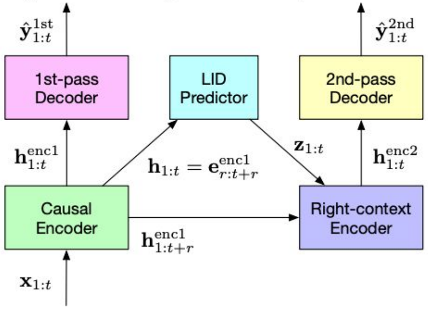
USM Results across ASR and ST tasks

Task	Multilingual Long-form ASR				Multidomain en-US	Multilingual ASR		AST
Dataset	YouTube			CORAAL	SpeechStew	FLEURS		CoVoST 2
Langauges	en-US	18	73	en-US	en-US	62	102	21
Prior Work (single model)								
Whisper-longform	17.7	27.8	-	23.9	12.8			
Whisper-shortform[†]	-	-	-	13.2[‡]	11.5	36.6	-	29.1
Our Work (single model)								
USM-LAS	14.4	19.0	29.8	**11.2**	**10.5**	**12.5**	-	-
USM-CTC	**13.7**	**18.7**	**26.7**	12.1	10.8	15.5	-	-
Prior Work (in-domain fine-tuning)								
BigSSL [3]	14.8	-	-	-	7.5	-	-	-
Maestro [67]					7.2			25.2
Maestro-U [67]							26.0 (8.7)	
Our Work (in-domain fine-tuning)								
USM	13.2	-	-	-	7.4	13.5	19.2 (6.9)	28.7
USM-M	**12.5**	-	-	-	**7.0**	**11.8**	**17.4 (6.5)**	**30.7**
Our Work (frozen encoder)								
USM-M-adapter[§]	-	-	-	-	7.5	12.4	17.6 (6.7)	29.6

Inferring Language ID with ASR



(a) *Predicted LIDs fed into the 2nd-pass decoder.*



(b) *Predicted LIDs fed into the right-context encoder.*

Figure 1: *Two ways to incorporate the frame-synchronous LID predictor into RNN-T with cascaded encoders.*

- Frame-synchronous LID predictor can provide streaming LID predictions at every frame
  - Used by the encoder and frame-synchronous decoder of the streaming RNN-T model.
  - Long right-context of the 2nd-pass decoding of cascaded encoders is suitable for predicting LIDs

Zhang, C., Li, B., Sainath, T.N., Strohman, T., Mavandadi, S., Chang, S.Y. and Haghani, P., Google LLC, 2023. Streaming End-to-end Multilingual Speech Recognition with Joint Language Identification. U.S. Patent Application 18/188,632.

We have been aiming to use ALL available data

Given a language, we can find a subset if not all of:

- Untranscribed (found) speech
- Unspoken text
- Paired ASR data (in-the-wild)
- Paired TTS data

How can we build usable ASR, Speech Translation, TTS systems in 1000s of languages with this?

Cross-modality and Cross-lingual Knowledge Transfer

- Maestro-U (ASR with zero-transcribed speech)
  - Modality specific encoders feed a shared encoder.
  - Language specific adapters in the shared encoder.
  - Labeled speech for some languages
  - Only unpaired speech and unpaired text for some languages
  - **NO LEXICON or G2P - Unicode Byte inputs support performance even on unseen scripts**
- Virtuoso (TTS with zero-transcribed speech)
  - Similar approach but applied to TTS
  - Speech decoder (feature to spectrogram) doesn’t see any transcribed audio.
  - **NO LEXICON or G2P - graphemic to acoustic form can be learned directly without explicit intermediate phone labels**

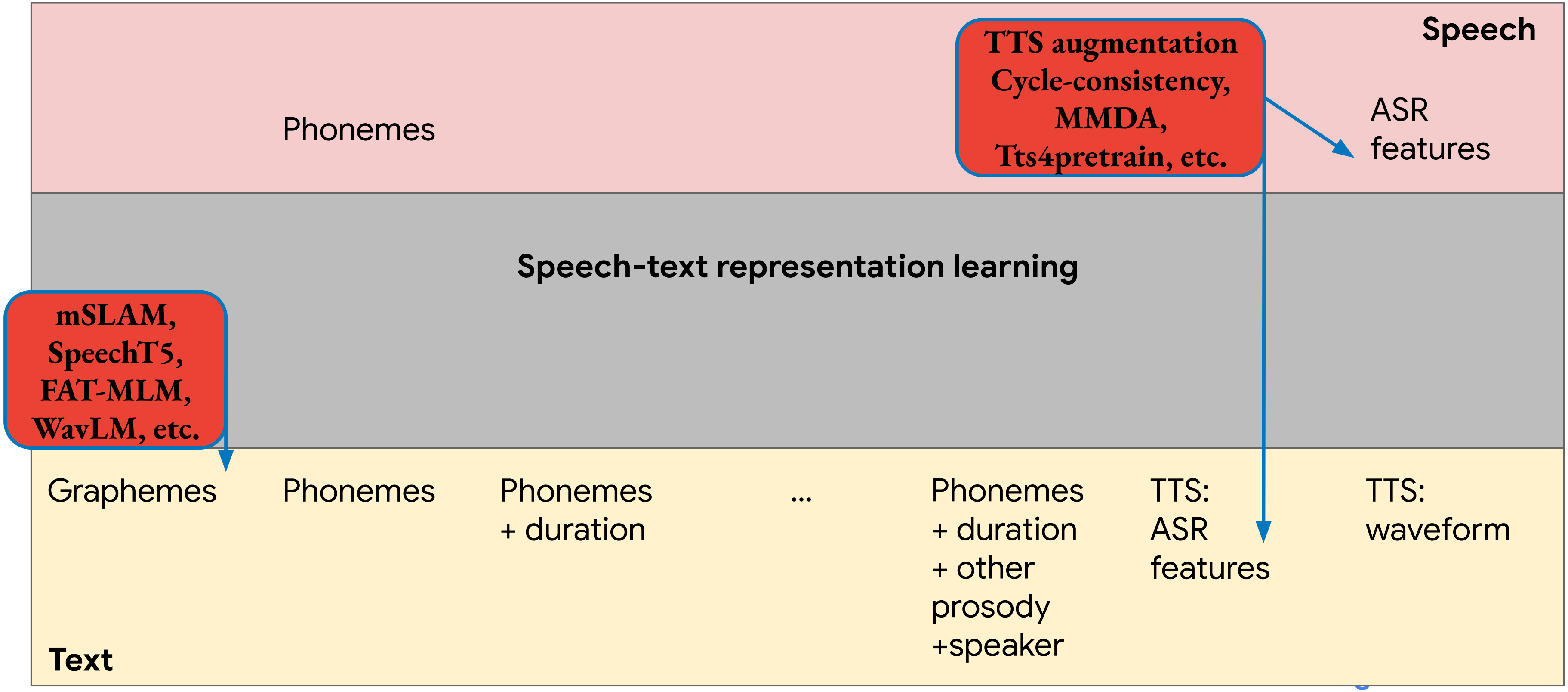
Google

Chen, Z., Zhang, Y., Rosenberg, A., Ramabhadran, B., Moreno, P., Bapna, A. and Zen, H., **MAESTRO: Matched Speech Text Representations through Modality Matching,** Interspeech 2022.

Speech-text representation learning

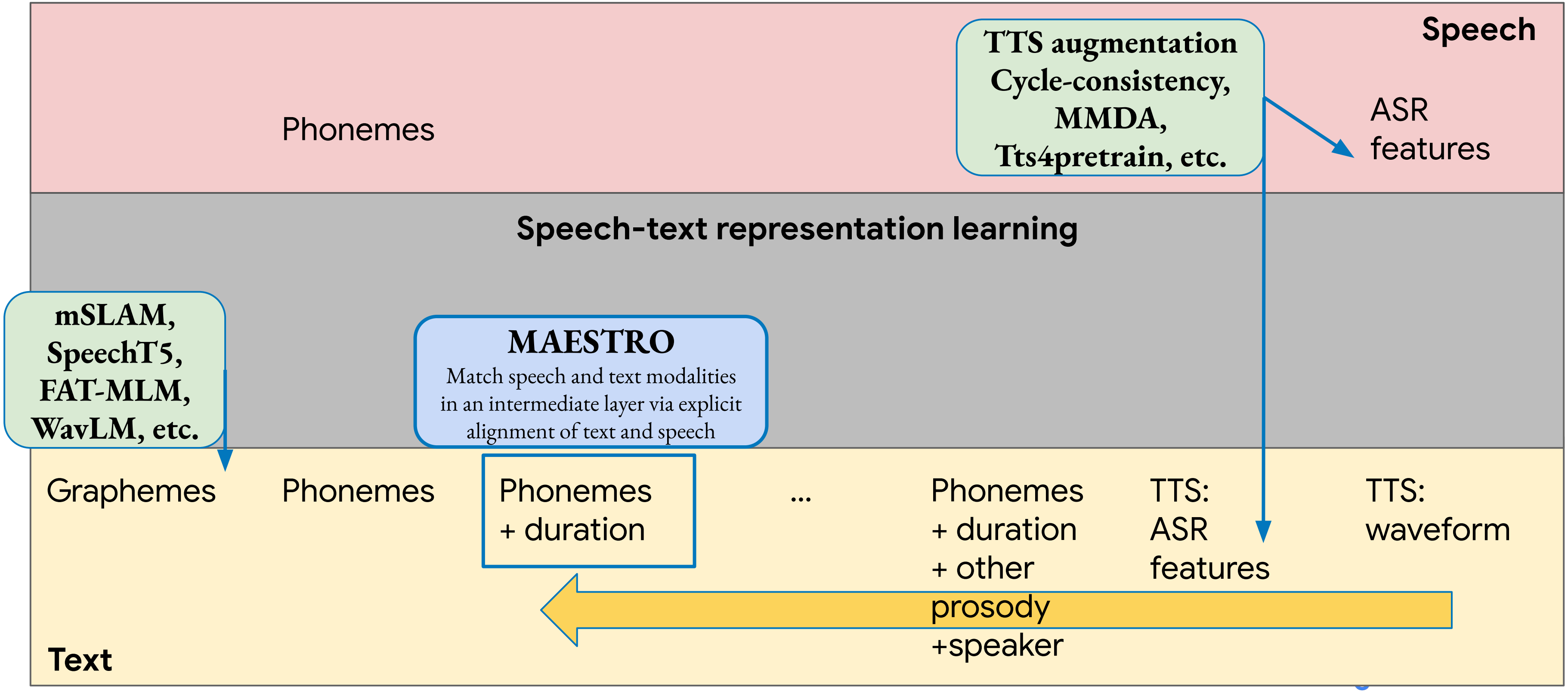
- **Complementary information** contained in Text and Speech[1]
  - **text**: domain; **speech**: acoustic conditions, speakers, etc.
- **Unify** speech and text representations
  - Simplify learning from both modalities
  - Learn better linguistic context in (conformer) encoders
- **Data minimization** by incorporating **unspoken text**
  - Low-resource speech processing

Joint speech+text representations



More related works can be referred to the paper.

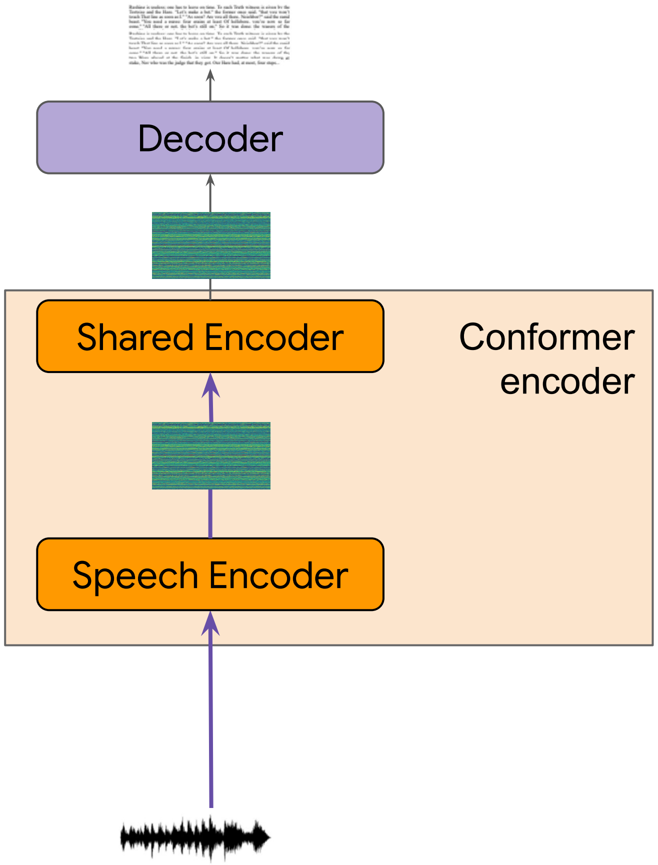
Joint speech+text representations



More related works can be referred to the paper.

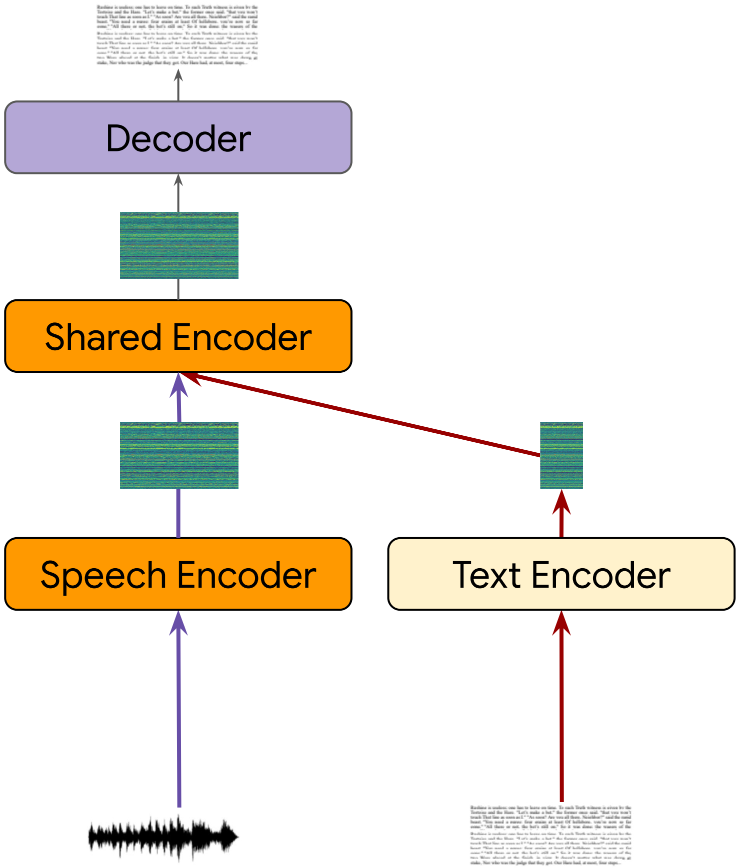
Architecture

Split original Encoder into two



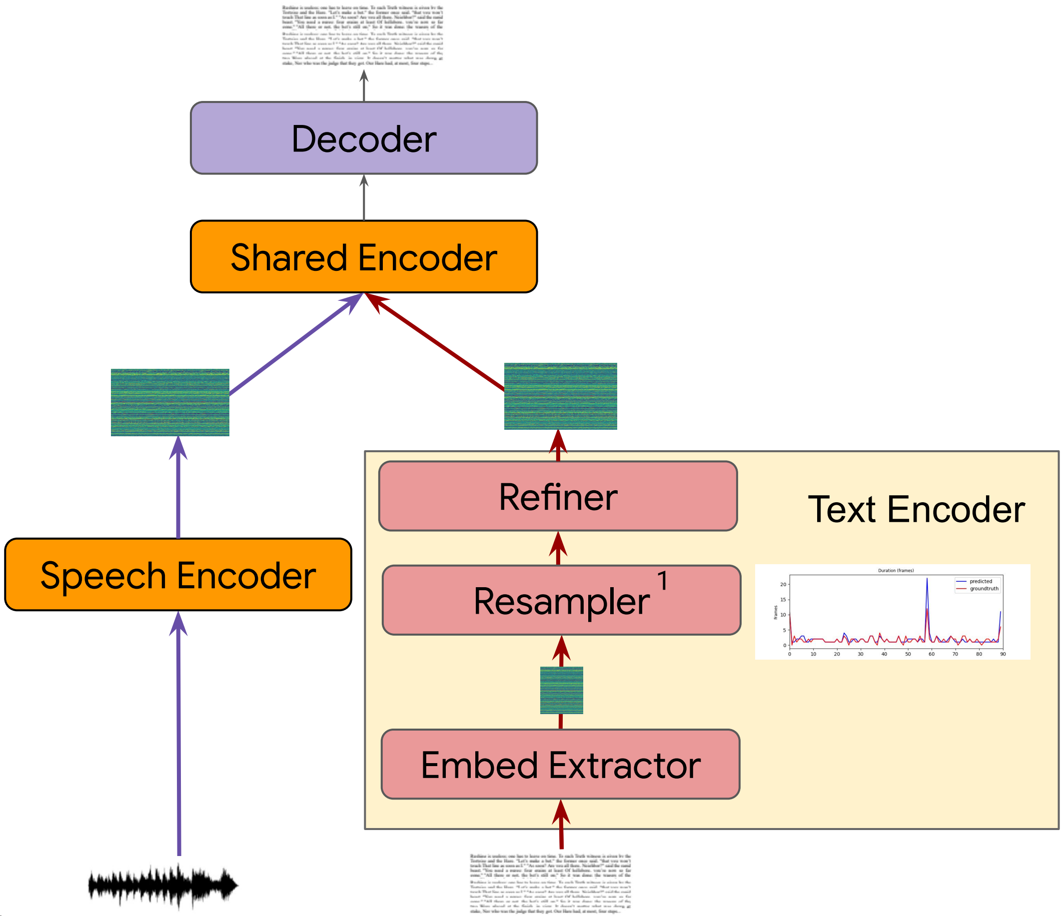
Architecture

Inject text representations in the middle



Architecture

How to match the two modalities?

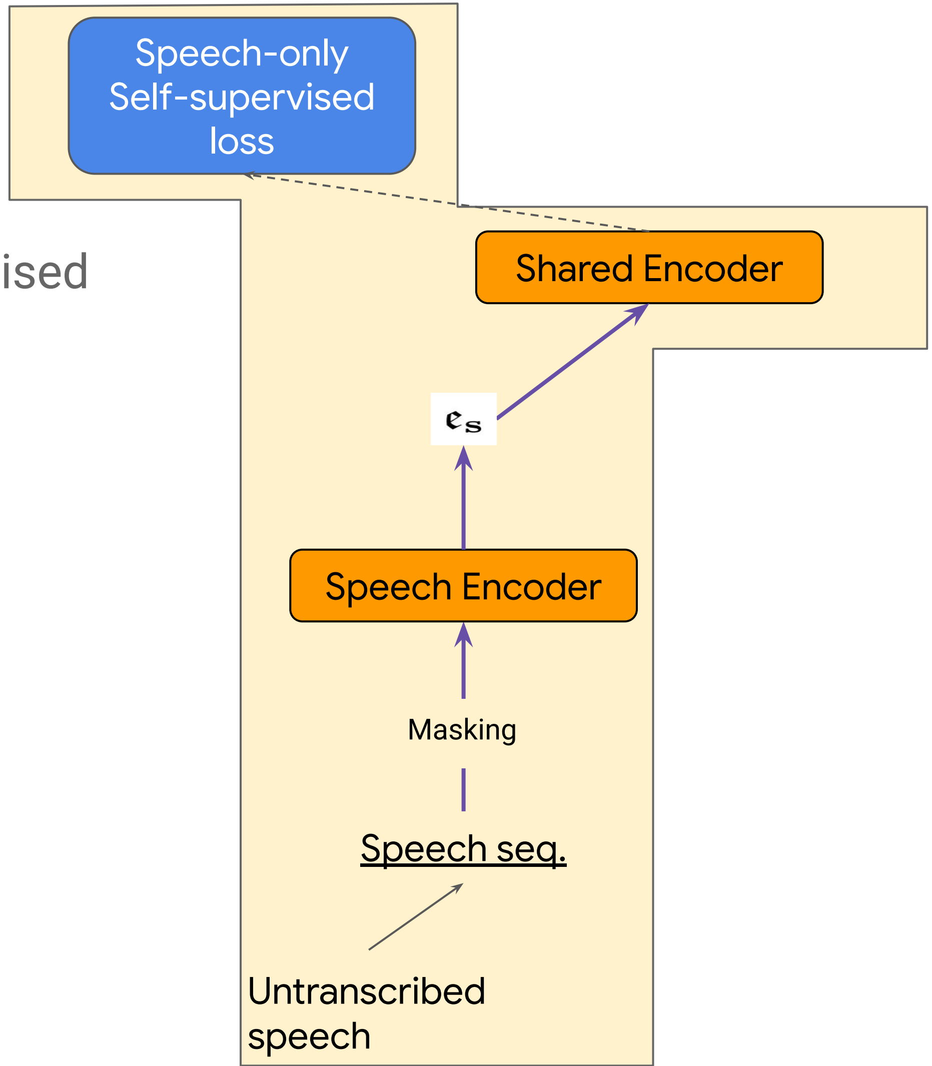


[1] Elias, Isaac, et al. "Parallel tacotron: Non-autoregressive and controllable tts." 2021.

Loss breakdown: **Speech-only**

Reuse any self-supervised pretraining objective

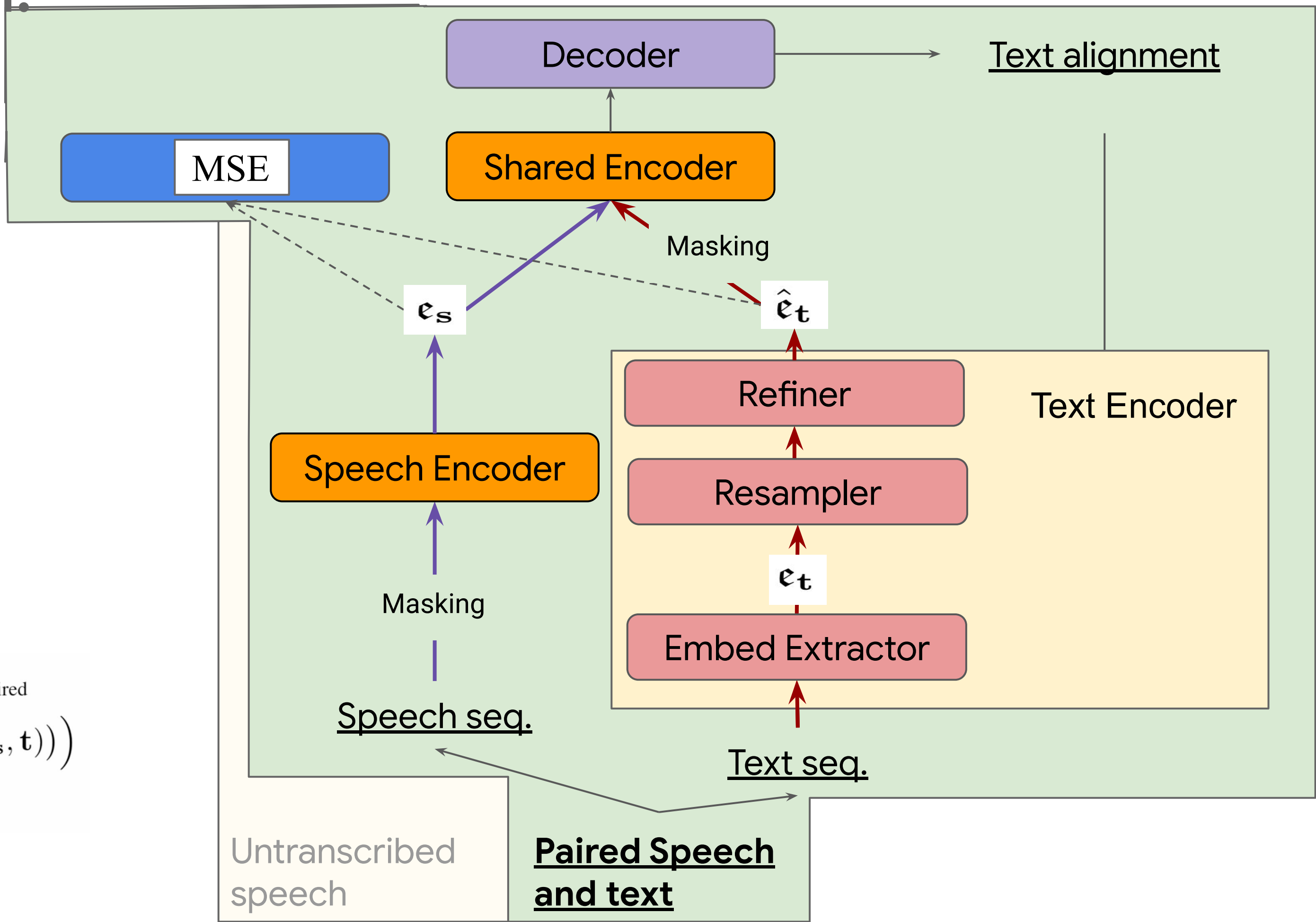
- W2v-BERT
- Best-RQ
- w2v1



Loss breakdown: **Paired Speech**

Train with $\mathcal{L}_{\text{MM}}$:

1. Align
2. Resample
3. Refine



$$\mathfrak{e}_{\mathbf{s}} = \theta_s(\mathbf{s}),\ \mathfrak{e}_{\mathbf{t}} = \theta_t(\mathbf{t}), \quad (\mathbf{t}, \mathbf{s}) \in \mathcal{X}_{\text{paired}}$$

$$\hat{\mathfrak{e}}_{\mathbf{t}} = \theta_{\text{Refiner}}\Big(\text{Resample}\big(\mathfrak{e}_{\mathbf{t}}, \text{Align}_{\text{Rnnt}}(\mathfrak{e}_{\mathbf{s}}, \mathbf{t})\big)\Big)$$

$$\mathcal{L}_{\text{MM}} = \text{MSE}(\mathfrak{e}_{\mathbf{s}}, \hat{\mathfrak{e}}_{\mathbf{t}}) + \mathcal{L}_{\text{Rnnt}}(\mathbf{t} \mid \mathfrak{e}_{\mathbf{s}})$$

Loss breakdown: Text-only

Inference using **Text Encoder**:

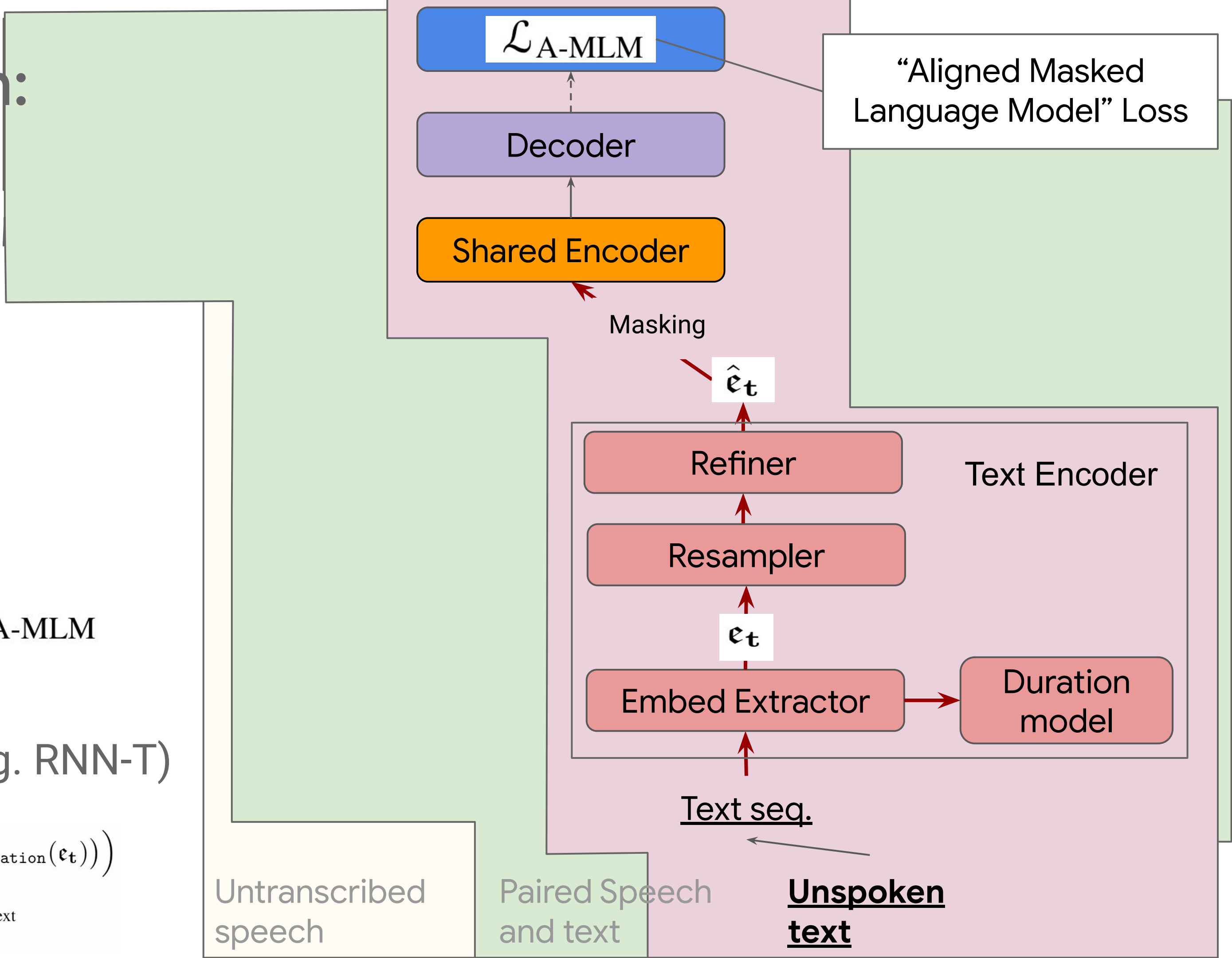
1. **Predict Duration**
2. Resample
3. Refine

Text learning with $\mathcal{L}_{\text{A-MLM}}$

1. Mask
2. Decoder loss (e.g. RNN-T)

$$\mathbf{e_t} = \theta_t(\mathbf{t}), \hat{\mathbf{e}}_\mathbf{t} = \theta_{\text{Refiner}}\Big(\text{Resample}\big(\mathbf{e_t}, \theta_{\text{Duration}}(\mathbf{e_t})\big)\Big)$$

$$\mathcal{L}_{\text{A-MLM}} = \mathcal{L}_{\text{Rnnt}}\Big(\mathbf{t} \mid \text{Mask}(\hat{\mathbf{e}}_\mathbf{t})\Big), \quad \mathbf{t} \in \mathcal{X}_{\text{text}}$$



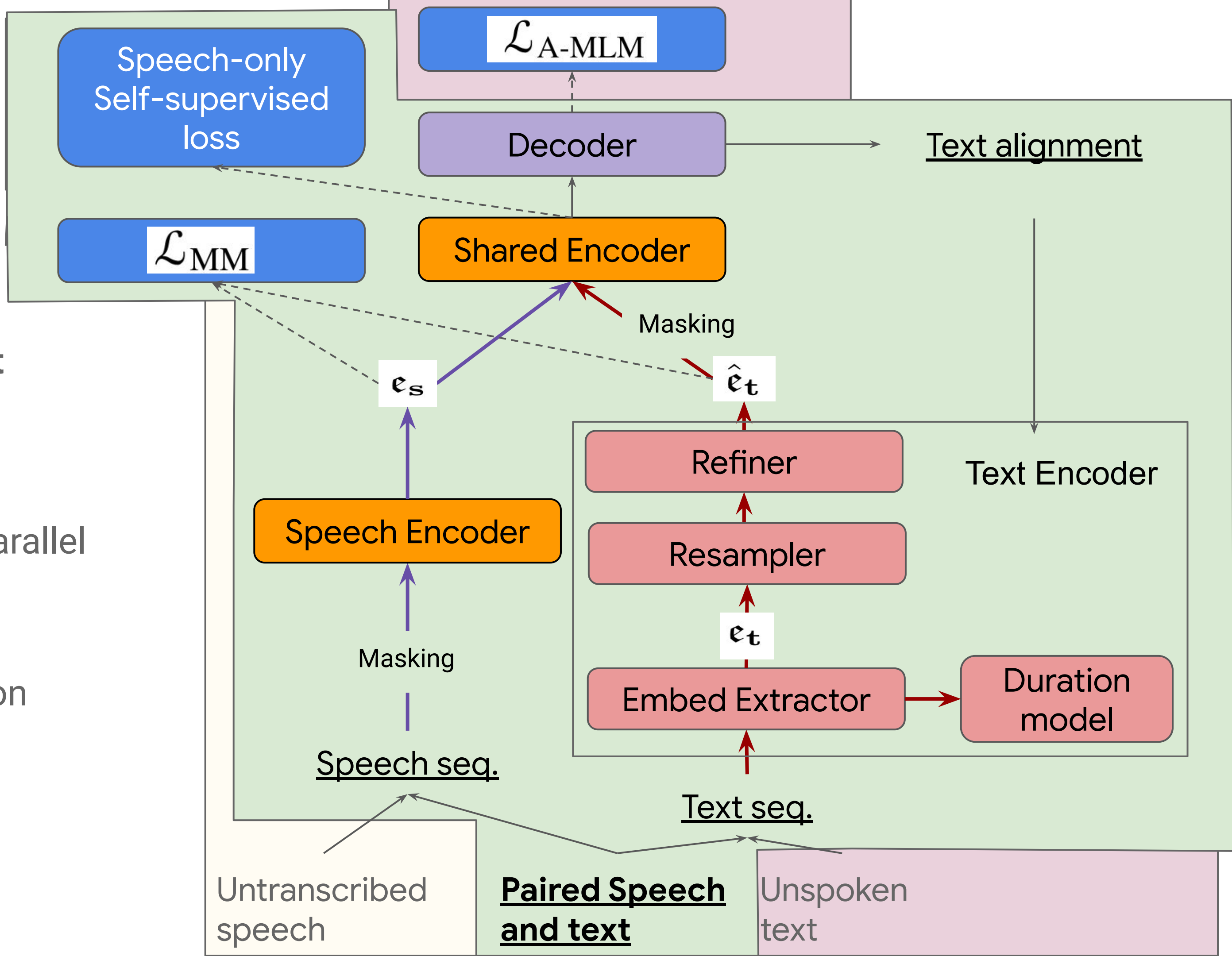
Overview

Sequence **self-alignment**

Modality matching in the intermediate layer

Reuse **duration** part of Parallel Tacotron

Unified framework for text-speech representation learning



Data and tasks

	Task	Lang	Speech (hours)	Text	Paired speech
SpeechStew	**Monolingual ASR**	1 (5 genres)	60k	6GB	5k
VoxPopuli	**Multilingual ASR**	14	430k	15TB	1.3k
CoVoST	**Speech-to-text Translation**	21	430k	15TB	2.9k

Multilingual ASR: Voxpopuli (14 languages)

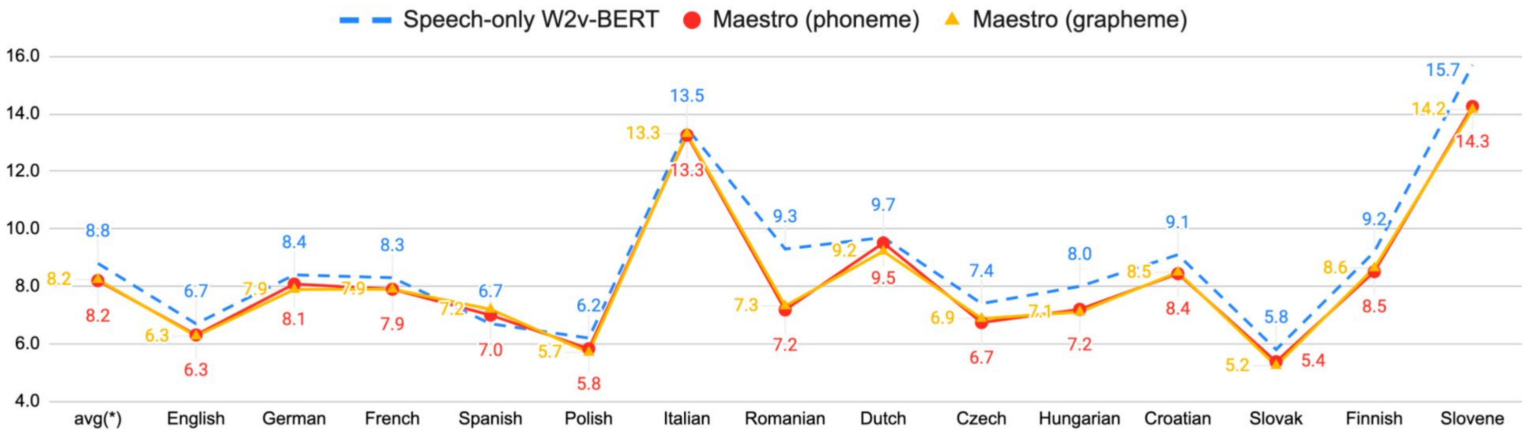
Method	Model size	Pretrain			Avg WER
		speech	text	paired data	
XLS-R	1B	437k	-	-	10.6
w2v-bert	0.6B	429k	-	-	8.8
Maestro	**0.6B**	**429k**	**VP-T + mC4**	**2.4k**	**8.1**

#1 New state-of-the-art
#2 Can be extended to cover 100 languages from mC4

Details in the submission.

Multilingual ASR: Voxpopuli (14 languages)

Breakdown: Languages are **sorted by the amount of paired data**



Generalize to different amount of paired data
No substantial difference from Phonemic and Graphemic modeling

Does this joint representation learning work on other tasks?

Speech-to-text Translation (ST, 21 languages->en)

Method	Model size	Pretraining Data					Avg BLEU
		Speech	Text	ASR	ST	MT	
Finetune: ST-only; mBART decoder init							
XLS-R	1B	437k	-	-	✗	✗	19.3
XLS-R	2B	437k	-	-	✗	✗	22.1
Finetune: ST and Machine translation (MT) jointly							
w2v-bert	0.6B	429k	-	-	✗	✗	21.0
mSLAM	0.6B	429k	mC4	2.4k	✗	✗	22.4
mSLAM	2B	429k	mC4	2.4k	✗	✗	24.8
Maestro	0.6B	429k	VP-T + mC4	2.4k	✗	✗	24.3
Maestro	**0.6B**	**429k**	**VP-T + mC4**	**2.4k**	✓	✓	**25.2**

Strong performance across ASR and Translation tasks

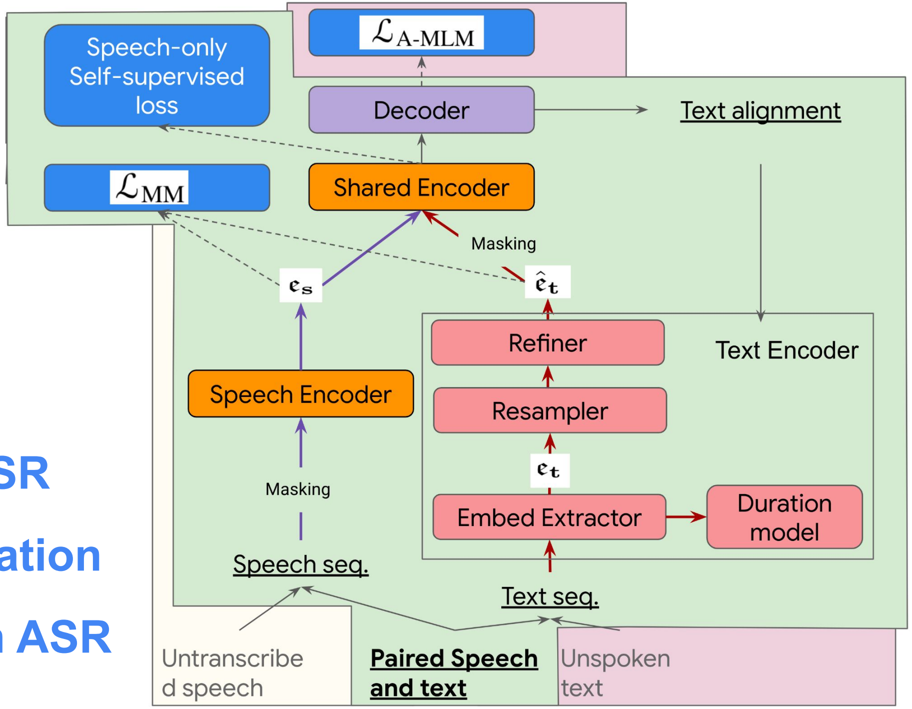
Numbers other than Maestro from "mSLAM: Massively multilingual joint pre-training for speech and text." link.

Key Finding:

Learn unified **speech-text** representations simultaneously that can transfer to diverse tasks

Solution: Maestro

- **Match speech and text modalities** in an intermediate layer via **explicit alignment of text and speech**
    - Sequence alignment
    - Matching modality embeddings
    - Duration prediction
    - Aligned masked-language model loss



Result: create new SOTAs

8% WER reduction on VoxPopuli **multilingual ASR**

2.8 BLEU improve on CoVoST 2 **Speech Translation**

4% WER reduction on SpeechStew **multidomain ASR**

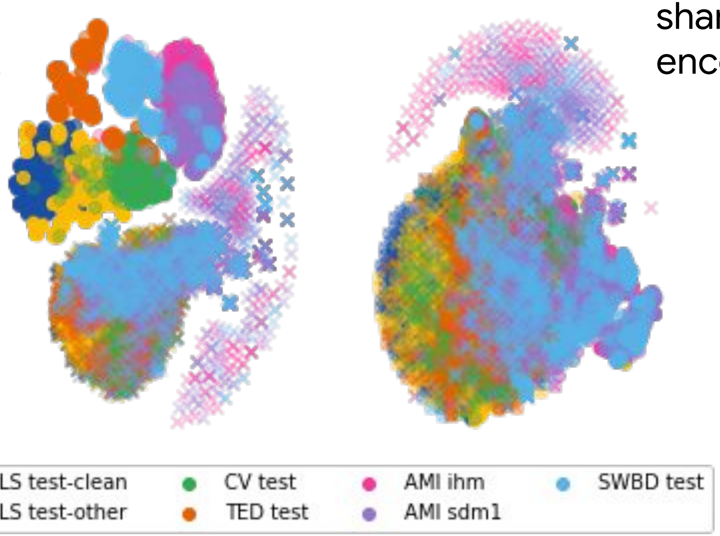
Retrieval to measure Shared Representation (ICASSP 2023)

Task: Given a speech sample, find the matching text sample or vice versa

unimodal encoders

Librispeech retrieval performance
test-clean: 20.5%
test-other: 19.3%

CV retrieval performance: 7.4%



shared encoder

Librispeech retrieval performance
test-clean: 83.5%
test-other: 68.8%

CV retrieval performance: 28.8%

Chance: 0.1% Other models at ~1-2%.
LibriSpeech trained encoders

Inspired by https://arxiv.org/abs/2209.15430 && https://arxiv.org/abs/2210.01738

Goal:

Train ASR **without transcribed speech** and **G2P**

Enable **multilingual transfer** even with unseen writing systems

Solution: Maestro-U

- Unsupervised speech and text learning with Maestro
- Promote multilingual knowledge transfer by Language ID and Residual Adapters
- Handling unseen writing systems by UTF-8 Bytes as text representation units

Result:

- Train ASR models without transcribed speech on 50 unseen FLEURS languages.
- Reduce the CER on languages with no supervised speech from 64.8% to 30.8%.
- Close the gap to oracle performance by 68.5% relative and reduces the CER of 19 languages below 15%.

Maestro-U: Leveraging joint speech-text representation learning for zero supervised speech ASR

Zhehuai Chen, Ankur Bapna, Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Pedro Moreno, Nanxin Chen

Handling unseen writing systems

Problem: How to share information across scripts?
When the text encoder, trained on seen languages, has never observed the script of unseen languages, the reliability of the alignments and thus the shared representation predictions suffer.

Solution: converting input graphemes (text) into a common representation that is “shared” across all languages.

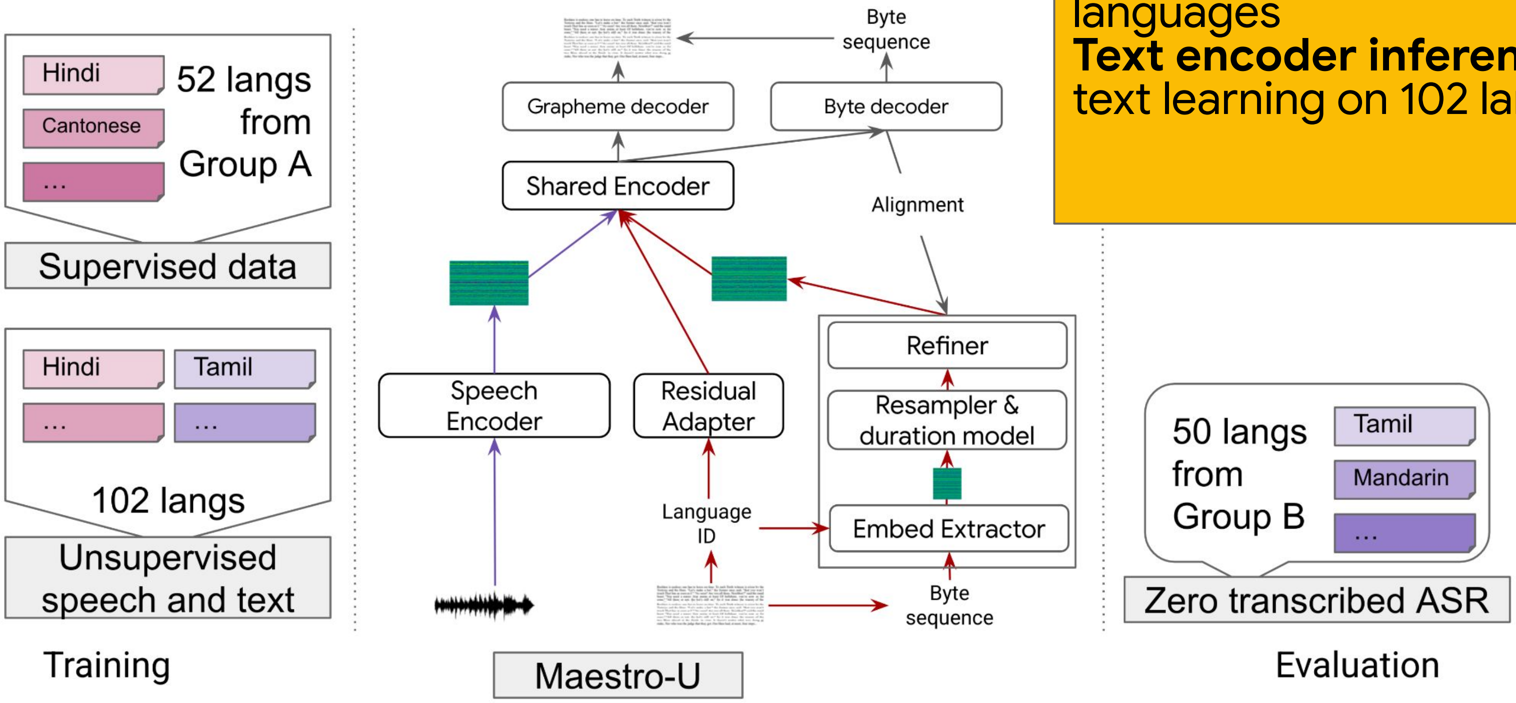
G2P and transliteration are higher resource solutions. Graphemic representations do not require additional resource of knowledge.
e.g. Bo Li et al 2019 “Bytes are all you need”; BPE from NLP.

No extra knowledge → Human knowledge

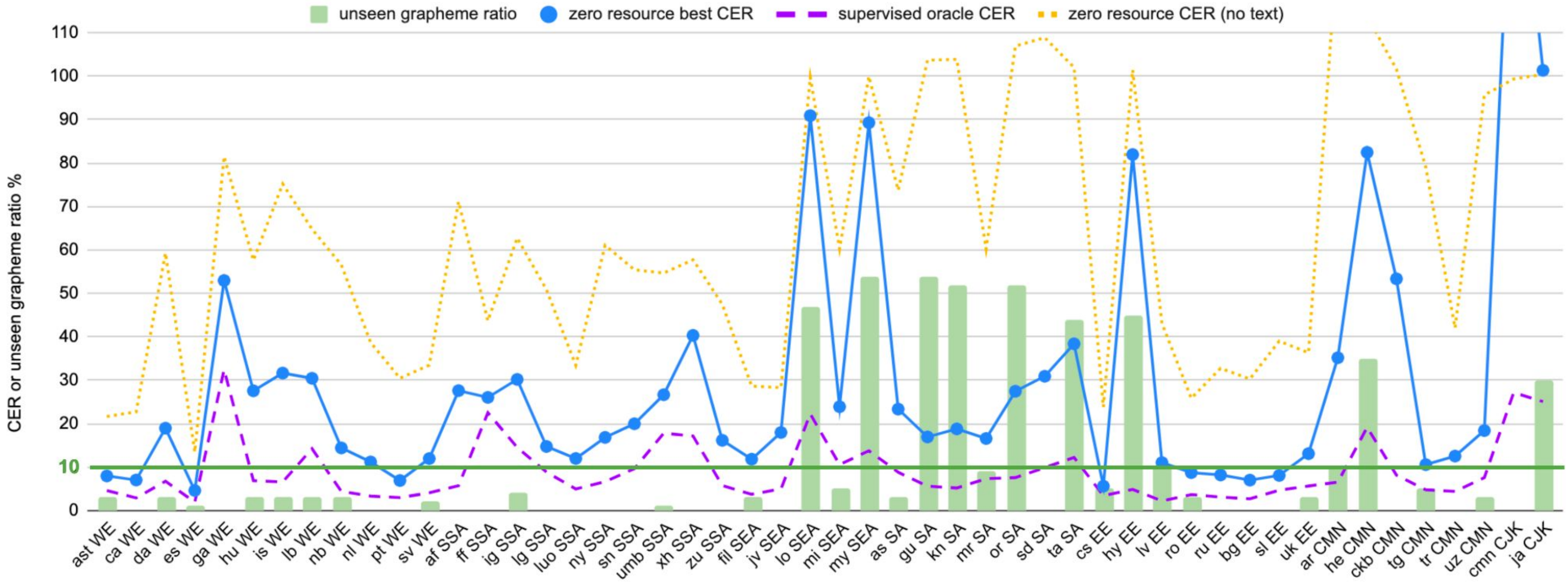
Grapheme > Byte-encoding → Transliteration → G2P

Massive multilingual ASR language expansion with zero supervised speech

Text encoder training: learn to predict speech-like text representations on 52 supervised languages
Text encoder inference: unspoken text learning on 102 languages



Results on 50 unseen languages (FLEURS)



Reduce the CER on languages with no supervised speech from 64.8% to 30.8%.
Even on the langs with very different writing systems, e.g. South Asian langs

Multilingual Text to Speech (TTS)

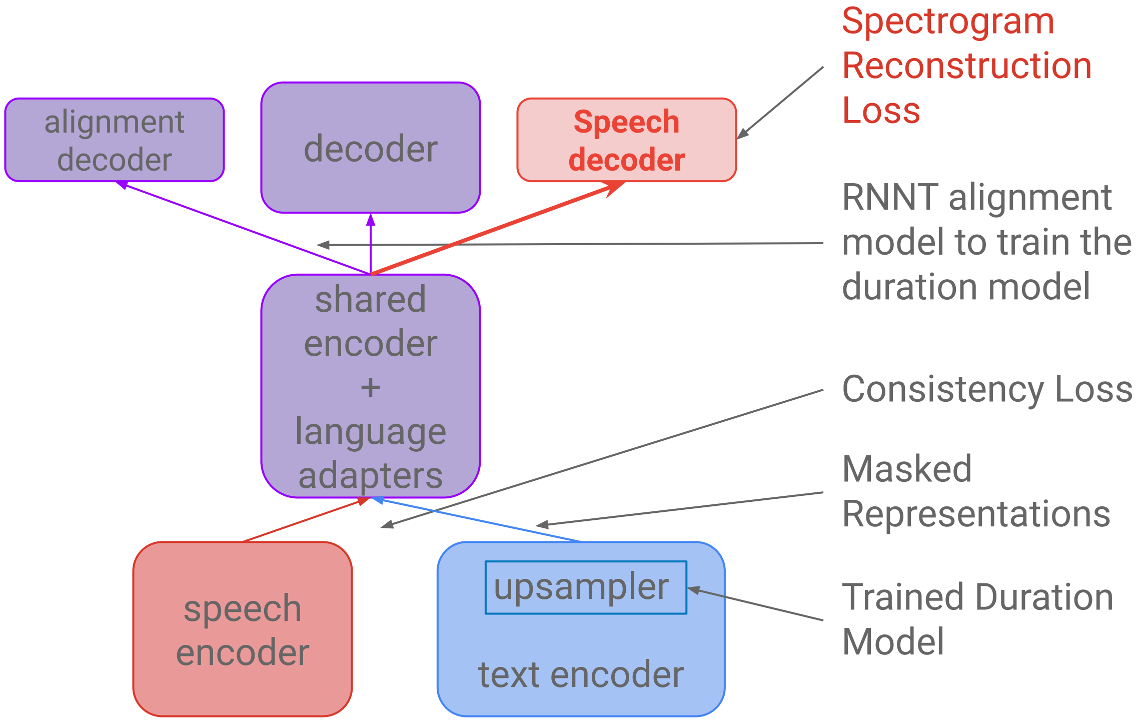
Current Google TTS covers around 60+ languages.

Around 7000 languages exist in the world.

Need to extend language coverage of TTS.

→ **Making use of various data to train multilingual TTS.**

Virtuoso = Maestro + speech decoder !!



Unpaired data ⇒ Self-supervised learning

- Sp enc → Sp dec ⇒ Masked AE
- Txt enc → Txt dec ⇒ Masked LM

Paired data ⇒ Supervised learning

- Text enc → Speech dec ⇒ TTS
- Speech enc → Text dec ⇒ ASR

Data

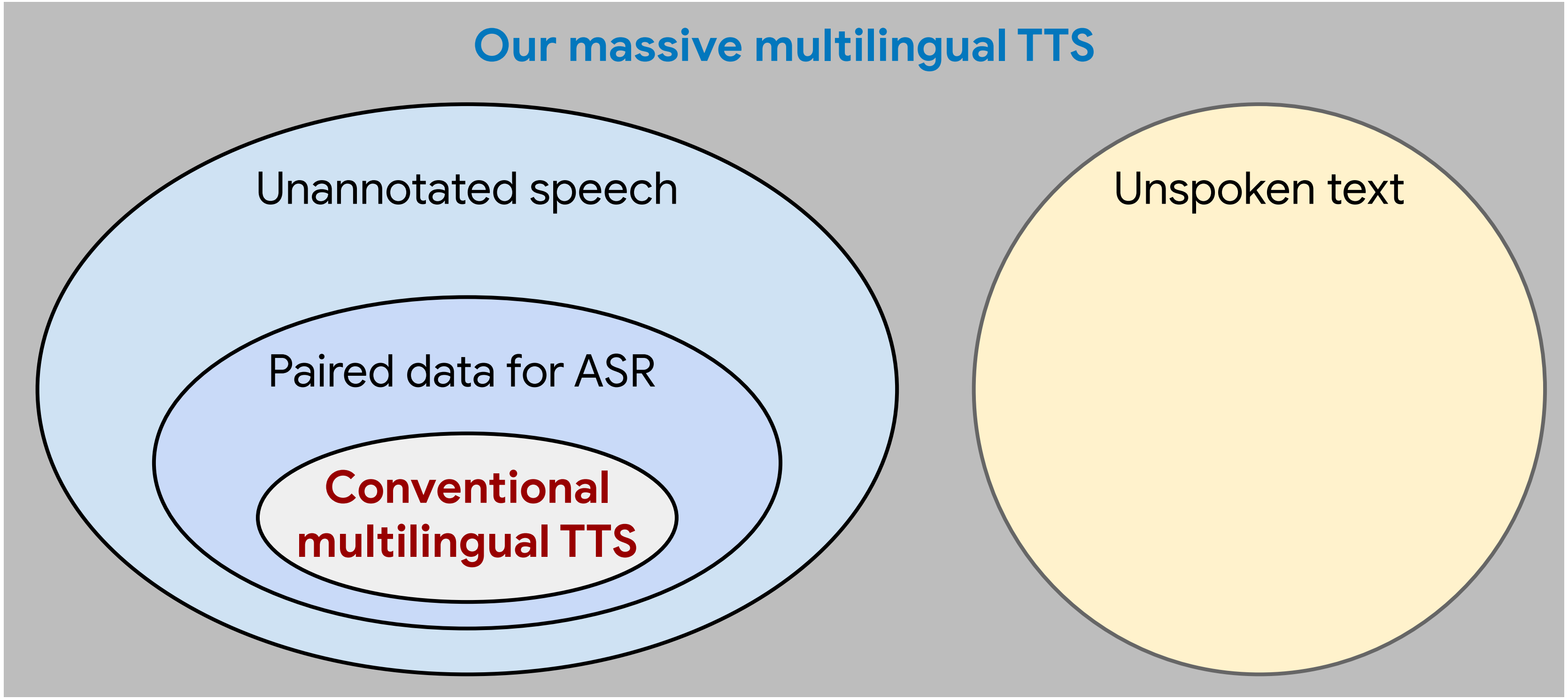
- Untranscribed speech
- Unspoken text
- Paired ASR data (in-the-wild)
- Paired TTS data (in-house)

Text representation

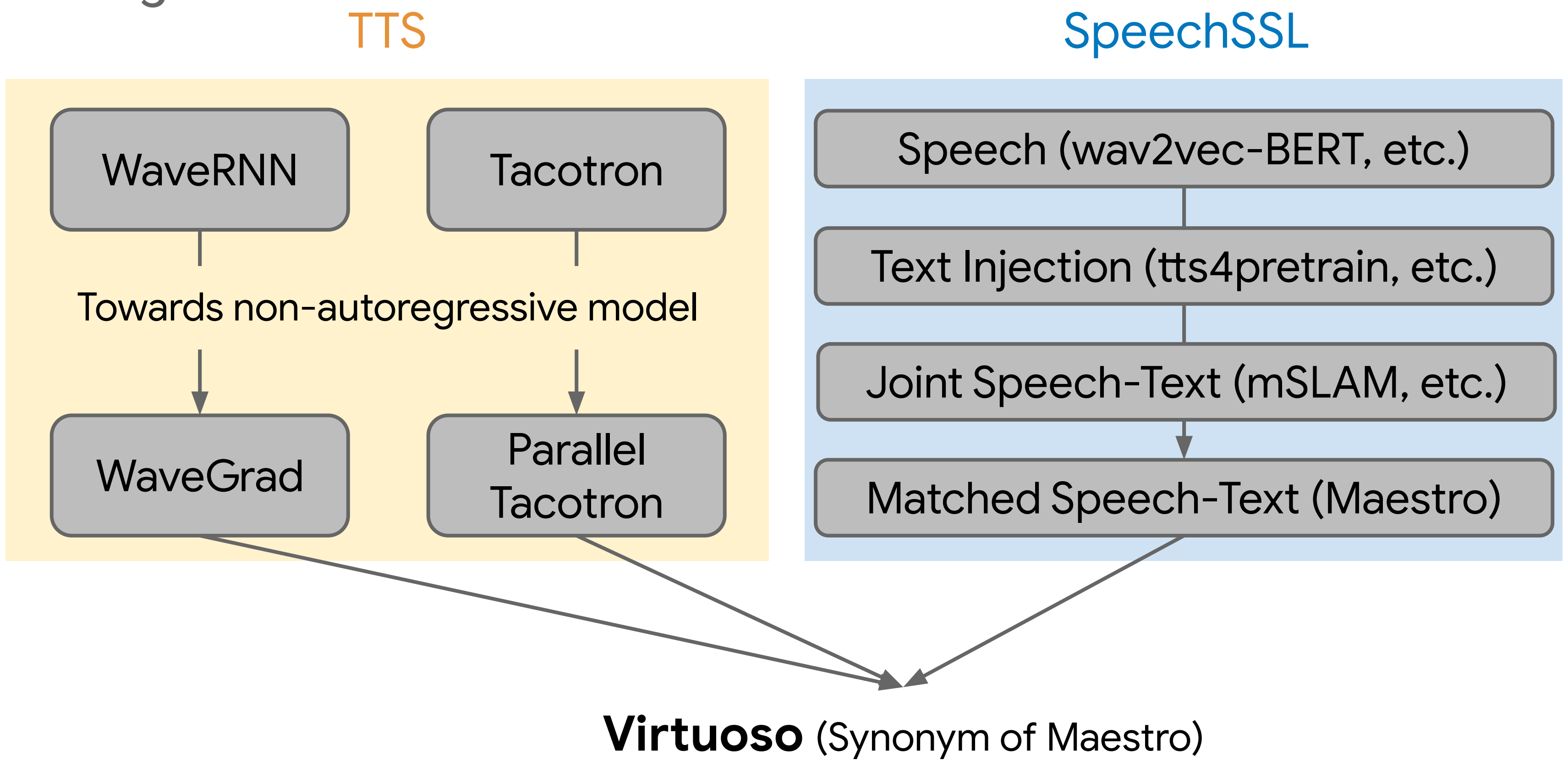
- Phonemes; Graphemes; Bytes

Training TTS model with massive unpaired speech text data

→ **Extending language coverages** without high data collection cost

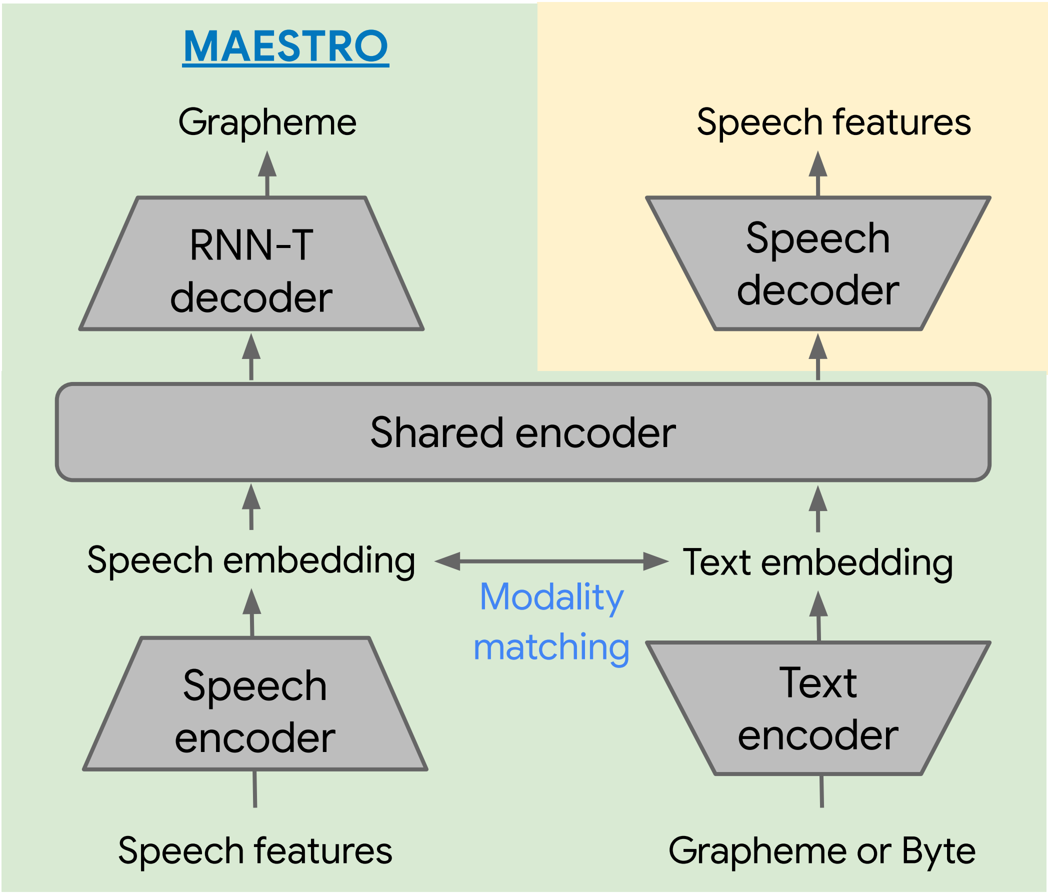


Meeting of Conventional TTS and Newer methods in ASR

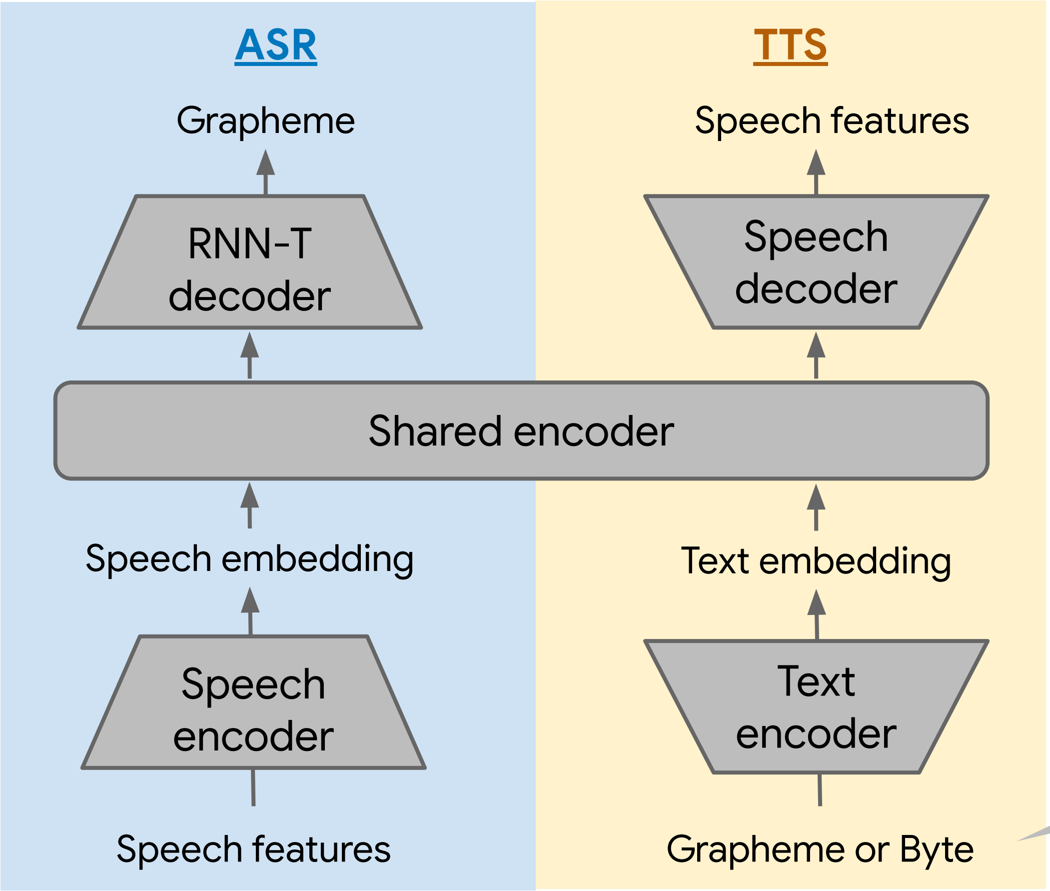


Overview of Virtuoso

Introducing speech decoder to speech-text SSL



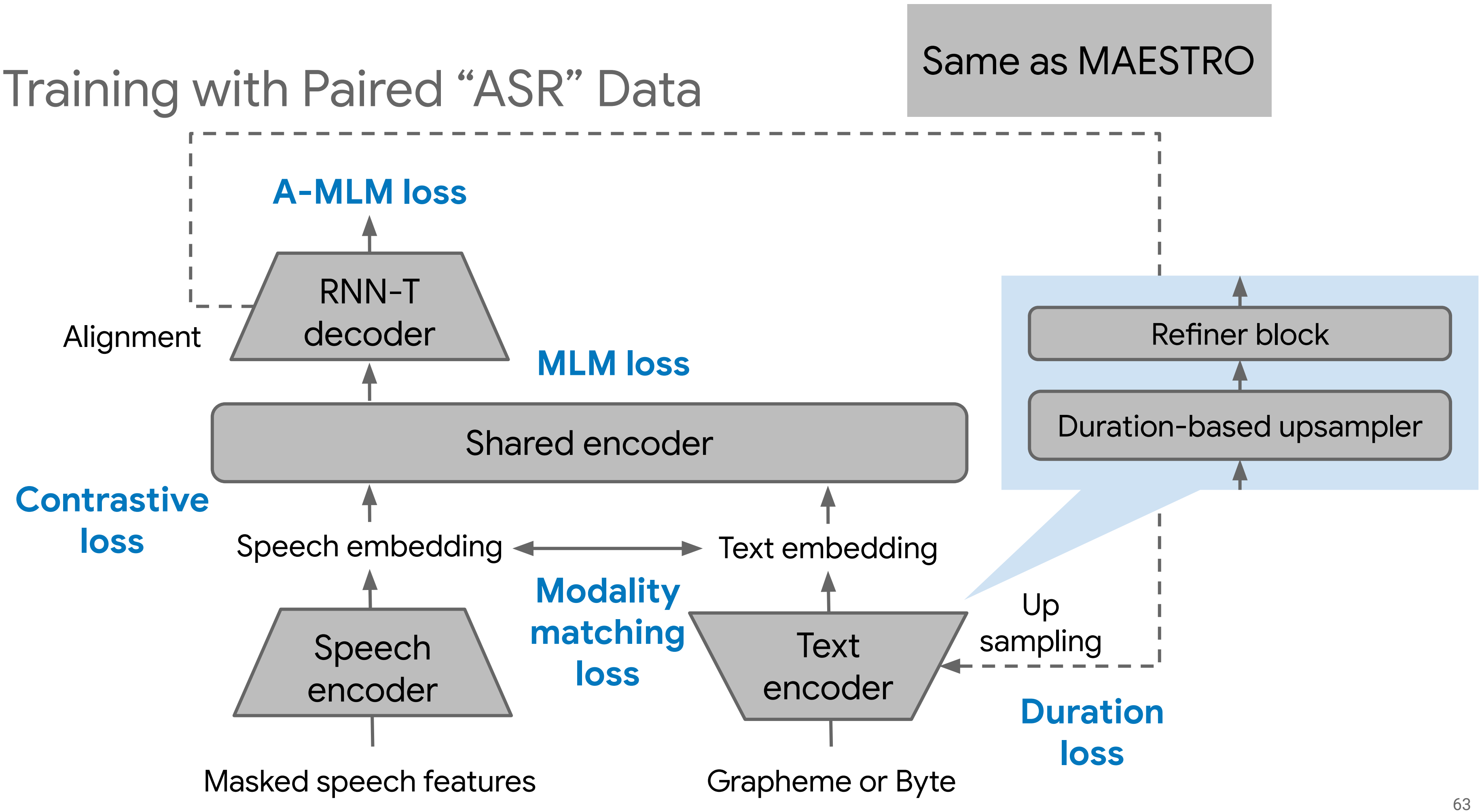
Consisting of ASR part and TTS part



We can obtain full TTS model without fine-tuning

Grapheme or Byte-based TTS without any G2P modules

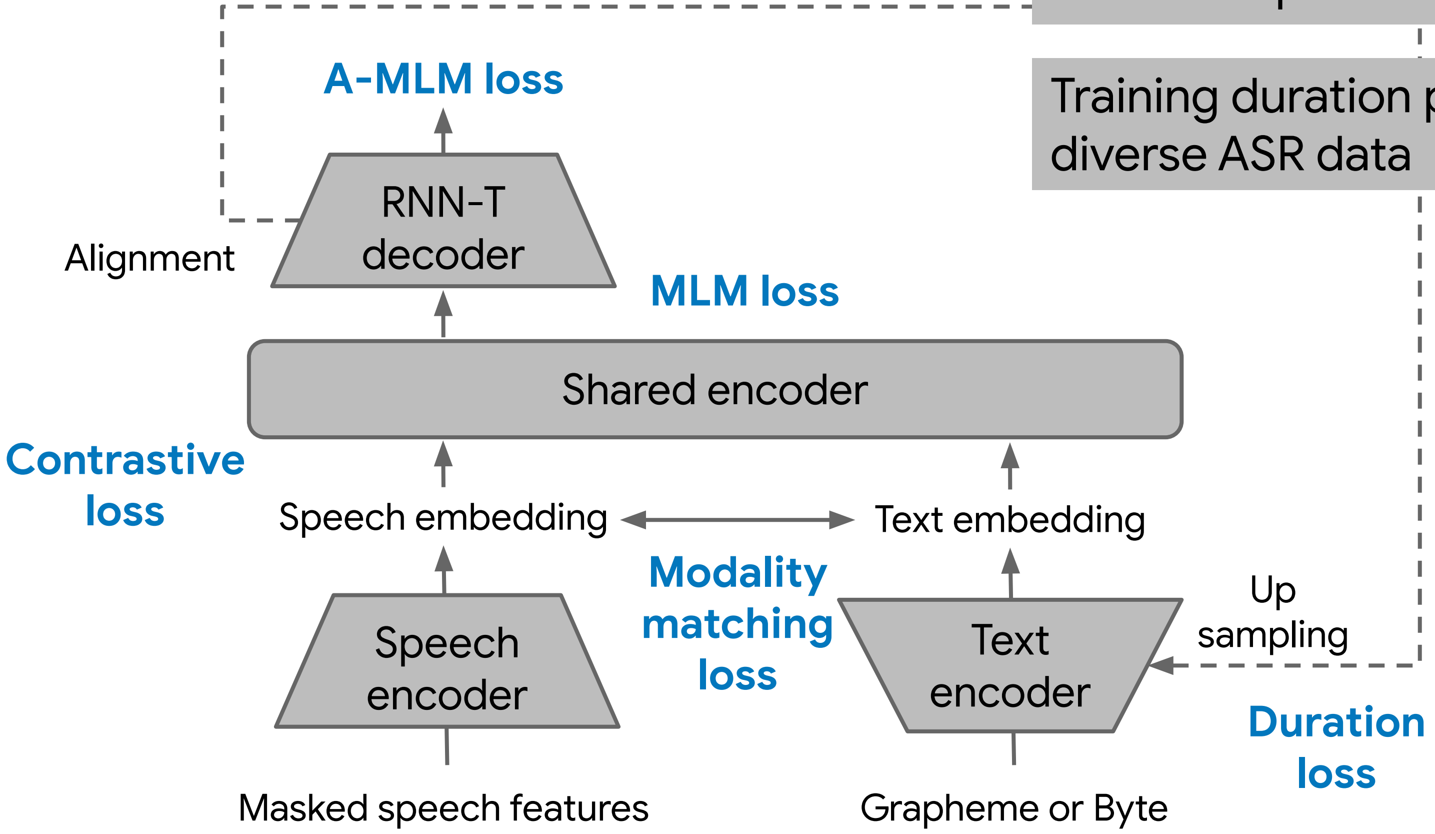
Training with Paired "ASR" Data



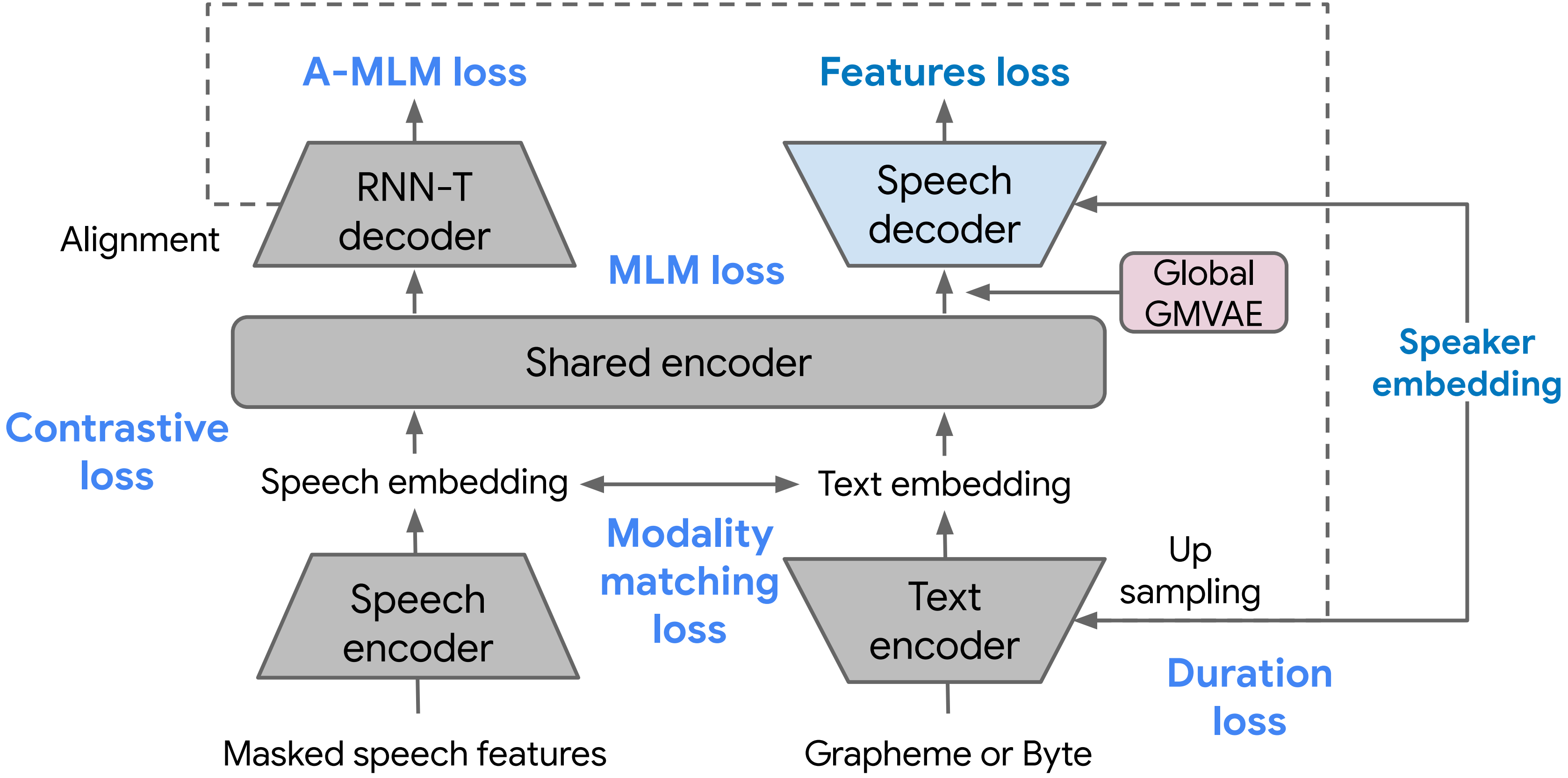
Training with Paired “ASR” Data

Learning shared representations between speech and text

Training duration predictor with diverse ASR data

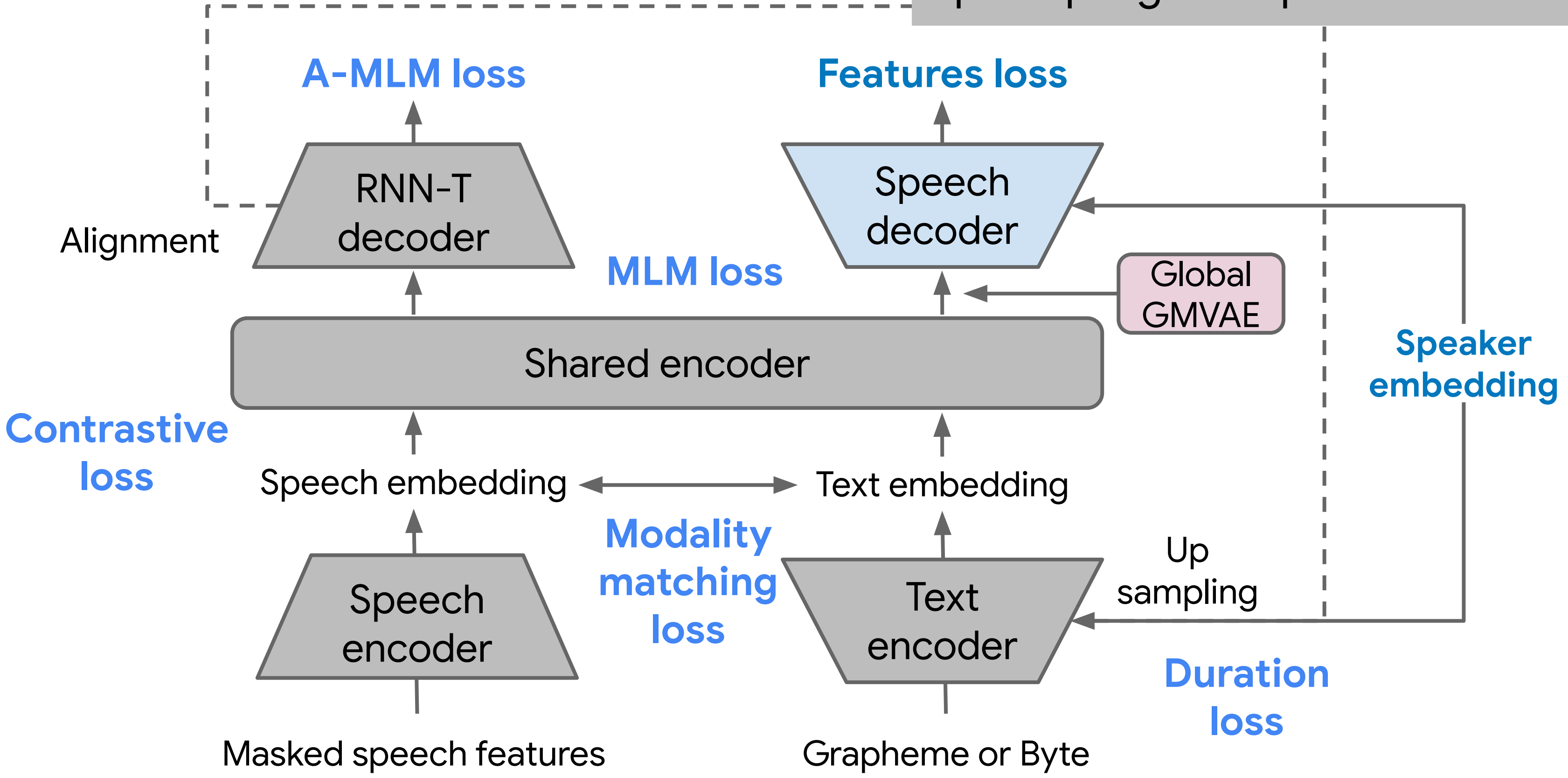


Training with Paired “TTS” Data



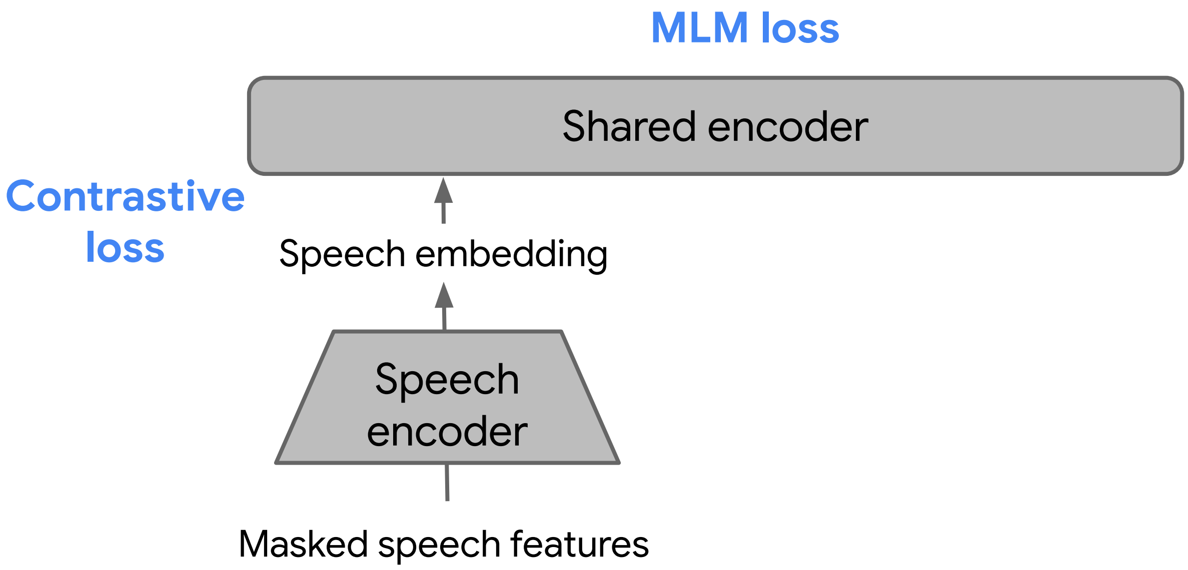
Training with Paired "TTS" Data

Injecting speaker embedding for text upsampling and speech decoding



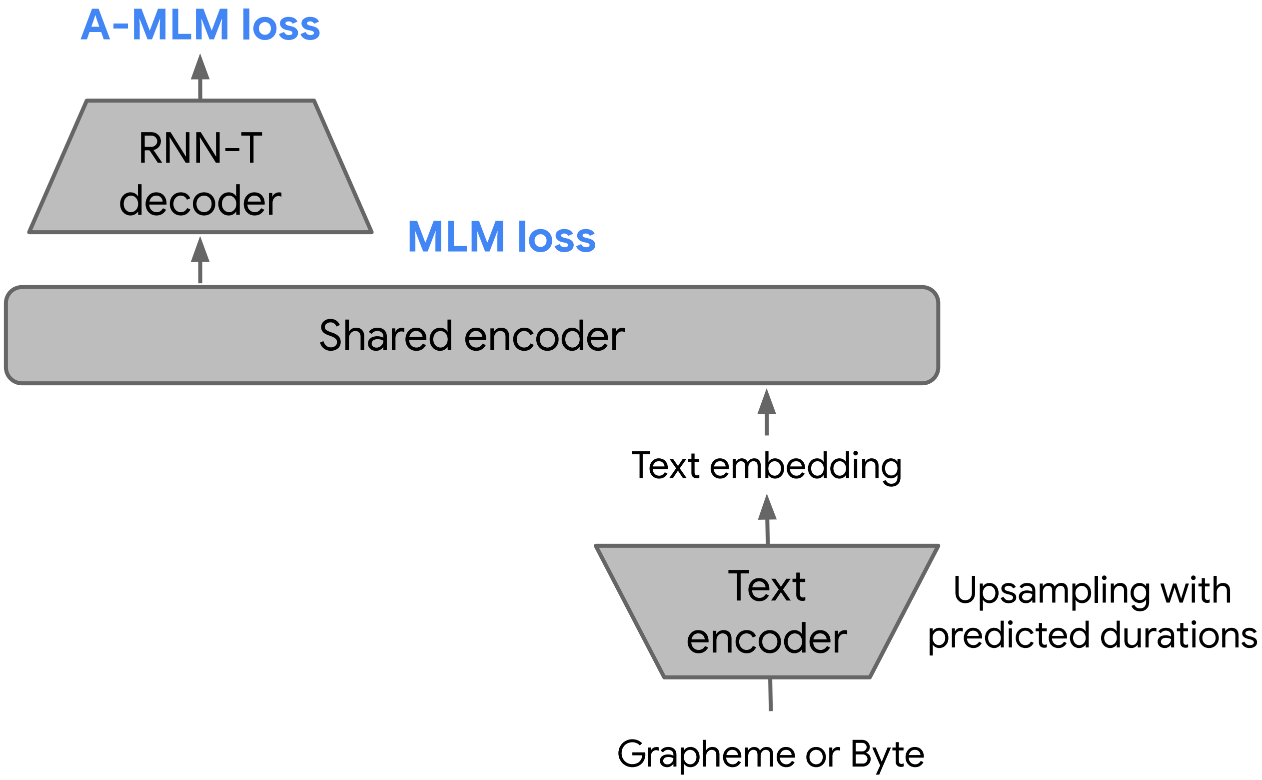
Training with Speech-Only Data

Same as w2v-BERT [Chung+21]

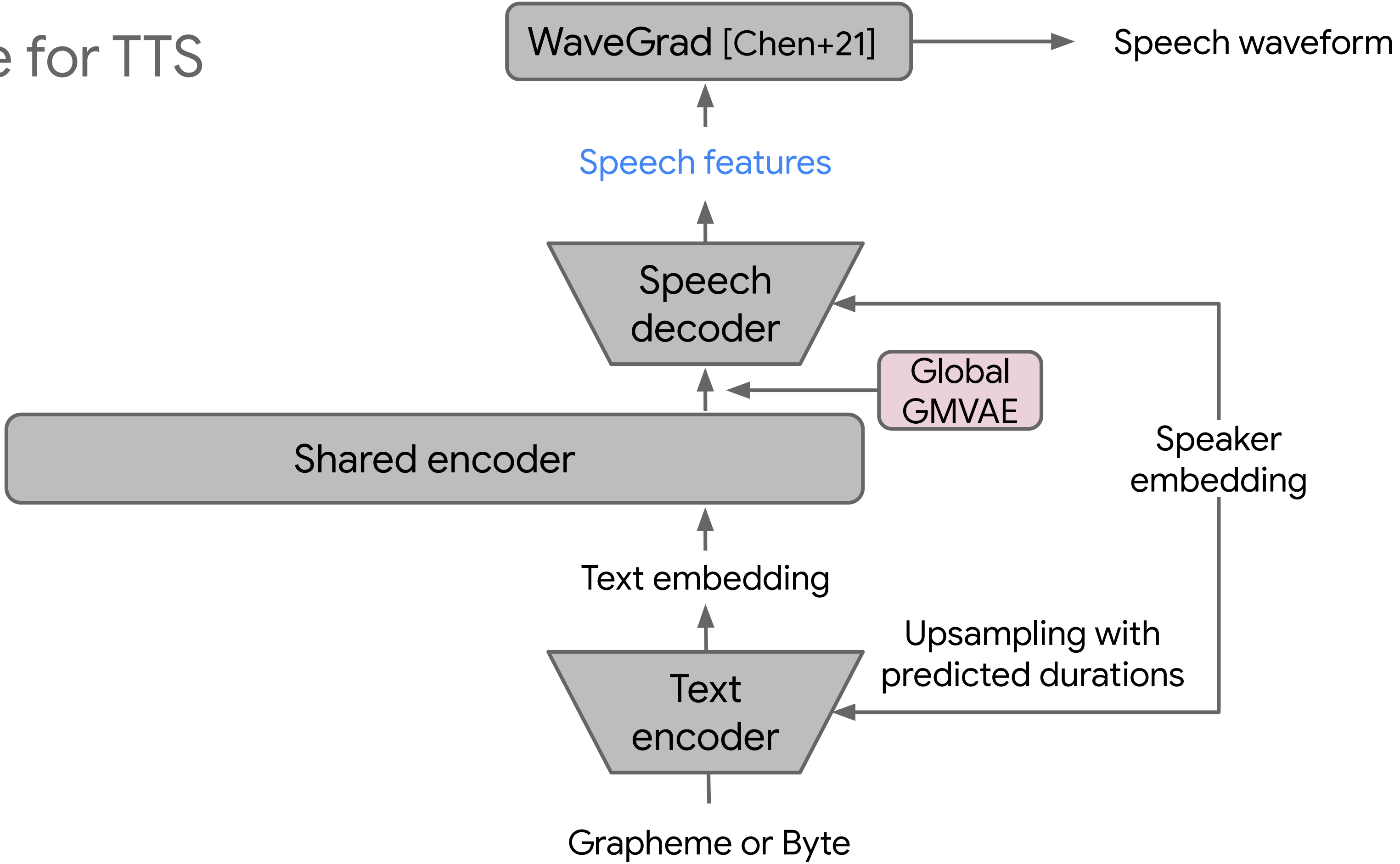


Training with Text-Only Data

Same as MAESTRO

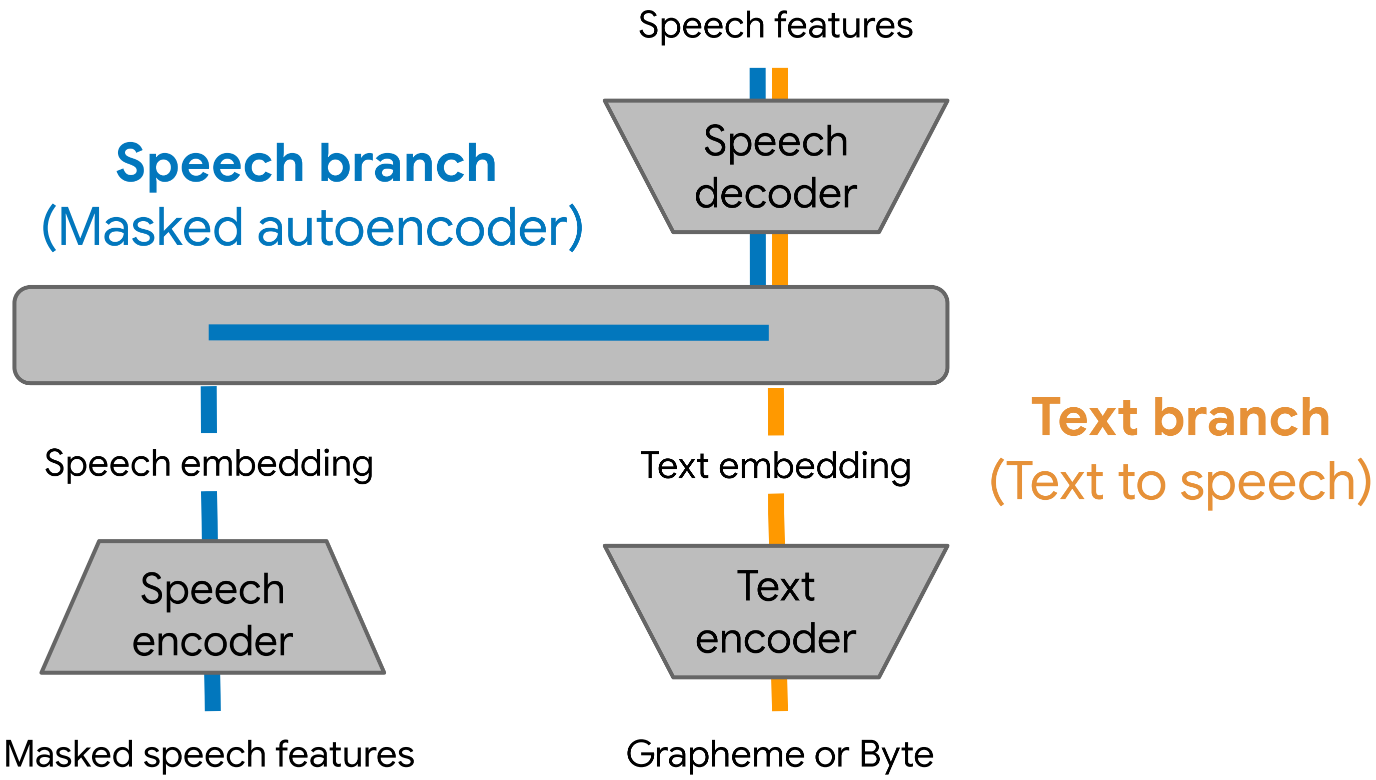


Inference for TTS



Random-Branch Training

Randomly switching speech branch and text branch to assist training of non-autoregressive speech decoder



Datasets

Paired TTS data	**40 languages, 1.5k h** PATTS: 44 locales
Paired ASR data	**96 languages, 3.3k h** Voxpopuli: 14 languages, 1.3k h MLS: 8 languages, 80h Babel: 17 languages, 1000h Fleurs: 96 languages, 960 h
Unpaired speech	**51 languages, 429k h** Voxpopuli, MLS, CommonVoice, and Babel
Unpaired text	**101 languages, 15TB** Voxpopuli: 3GB MC4: 101 languages, 15TB

Evaluation Metrics

1. **Mean opinion score (MOS):** Subjective test commonly used in TTS

   Evaluating **naturalness** of synthetic speech

2. **TER:** Token error rates calculated with a pretrained MMASR model

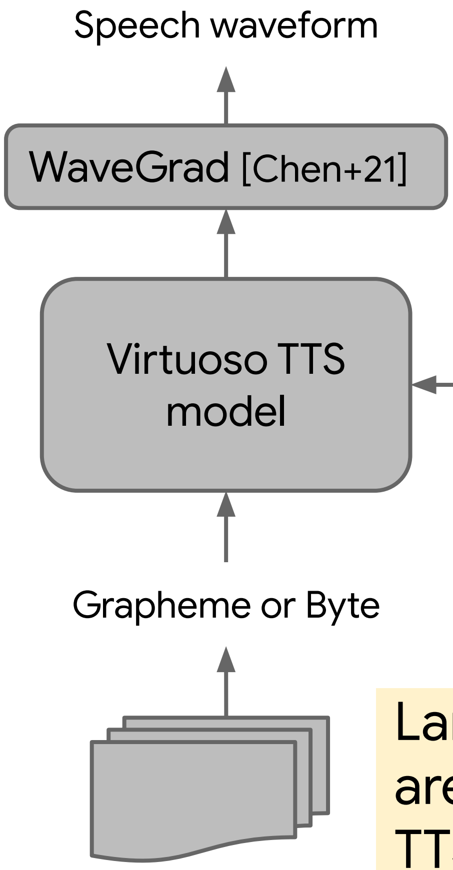
   Evaluating **accuracy of linguistic contents**

3. **SQuId:** Automatic MOS prediction model trained on 60 locales

   Evaluating **speech quality**

Zero-Resource TTS

Can massive multilingual knowledge obtained with ASR and SSL be transferred to TTS?



Speaker embedding

Sampled from similar locales included in TTS data

Languages which are not included in TTS training data

Evaluation of Low-Resource Locales

	Slovenian (0.3h)		**Farsi** (2.5h)	
	TER	SQuId	TER	SQuId
Natural	0.178	-	0.037	-
Tacotron2-G	0.109	3.87	0.045	3.41
Maestro-Finetune-G	0.139	3.87	0.056	3.66
Virtuoso-G-Paired	**<u>0.068</u>**	**<u>3.99</u>**	0.049	**<u>3.85</u>**
Virtuoso-G-All	0.073	3.93	**<u>0.044</u>**	3.77
Virtuoso-B-LID-All	0.070	3.92	0.069	3.82

In sl/si, larger gap in TER between baseline methods and Virtuoso

Virtuoso-G-Paired showed good results in Holdin locales

Evaluation of Zero-Resource languages

Nearest locales in the language family tree are **NOT** included.

	Tamil (0h)		**Turkish** (0h)	
	TER	SQuId	TER	SQuId
Natural	0.163	-	0.053	-
Tacotron2-G	0.928	3.39	0.748	3.74
Maestro-Finetune-G	0.952	2.62	0.819	3.99
Virtuoso-G-Paired	0.274	**4.35**	0.380	4.02
Virtuoso-G-All	**0.250**	4.23	0.241	**4.06**
Virtuoso-B-LID-All	0.295	4.15	**0.202**	4.03

Baseline methods did not work well.

Unpaired data significantly improved TER.

Demonstration of Low- and Zero-Resource Languages

	Slovenian (0.3h)	**Bulgarian** (0h)	**Tamil** (0h)
Natural			
Tacotron2-G			
Maestro-Finetune-G			
Virtuoso-G-TTS			
Virtuoso-G-Pair			
Virtuoso-G-All			
Virtuoso-G-Lid-All			
Virtuoso-B-Lid-All			

Fine-tuning on Zero-Resource Locales

Fine-tuning on zero-resource locales further improved TER.

Few-shot (1h) adaptation achieved decent performance.

	Tamil		Turkish		Bulgarian	
	TER	SQuId	TER	SQuId	TER	SQuId
Natural	0.163	-	0.053	-	0.052	-
Zero-Resource	0.250	4.23	0.241	**4.06**	0.256	3.83
Few-shot (1h) Fine-tuning	**0.187**	**4.28**	0.083	3.94	0.110	4.06
All-data Fine-tuning	0.211	4.15	**0.064**	3.97	**0.076**	**4.10**

Results of Subjective Evaluations

Virtuoso showed higher MOS than baseline methods

Virtuoso showed 3.39 MOS even for a zero-resource language

	English	French	Spanish	Tamil
Tacotron2-G	3.31±0.045	3.60±0.068	3.53±0.085	1.59±0.088
Maestro-Finetune-G	3.67±0.040	3.85±0.060	3.66±0.070	1.24±0.051
Virtuoso-G-TTS	1.87±0.050	2.35±0.109	1.60±0.095	1.28±0.069
Virtuoso-G-Paired	3.79±0.041	3.95±0.059	3.96±0.069	3.39±0.083
Virtuoso-G-All	3.81±0.039	3.86±0.065	3.89±0.074	2.98±0.078
Virtuoso-G-LID-All	1.89±0.037	2.14±0.087	2.36±0.078	1.89±0.077
Virtuoso-B-LID-All	3.71±0.041	3.82±0.066	4.01±0.065	2.89±0.083

Multilingual TTS possible with the same ASR technology

- Virtuoso improved performance for **both major and low-resource locales.**
- Virtuoso performed well in **zero-resource settings**.
- **Byte-based model** achieved the highest linguistic accuracy.
- **Only using paired ASR+TTS** data was better in terms of naturalness.
- **Using unpaired data** was effective for zero-resource settings.

Takaaki Saeki et al., , EXTENDING MULTILINGUAL SPEECH SYNTHESIS TO 100+ LANGUAGES WITHOUT TRANSCRIBED DATA, ICASSP 2024

Representation Learning

Learning Within Modality

Audio: Full-sum/Sampling based Distillation, Sampling Guided-masking, Diffusion-based masking, Use of ephemeral sources (eg.Radio/Podacsts), SoundStream + AudioLM

Text: Large LMs integrated into e2e model

Learning Across Modalities

Encourage unified representations

Share language adapters within language families

Acoustic Prompting

Additional modalities/signals (image, video, tonal language, etc.)

Intermediate representations help other downstream tasks (phone recognition, NLP?)

Weak Supervision

Conditional adapters (on topic, contextual keywords)

Grounding around other information seen in the same context (text/audio/image/audio)

Concluding Remarks

- Code-switching is by no means a solved problem for ASR or other ST/TTS tasks
- Well-represented Data Resources are scarce
- Language Identification is crucial and still remains a difficult problem for several code-switched languages
- Joint speech-text representation learning is useful for ASR, ST, TTS....
- For Indic Languages there is work underway via Bhashini (Natural Language Translation Mission)

Machine Learning continues to produce large models that can scale and be prompted to solve these tasks. These fundamental challenges remain and more research in these areas will pave the way for usable, scalable, multilingual models.